SOUTHWEST FISHERIES SCIENCE CENTER

NATIONAL MARINE FISHERIES SERVICE SOUTHWEST FISHERIES SCIENCE CENTER P.O. BOX 271 LA JOLLA, CA 92038

DECEMBER 1991

CALIBRATION OF SHIPBOARD ESTIMATES OF DOLPHIN SCHOOL SIZE FROM AERIAL PHOTOGRAPHS

By

LIBRARY

JAN 30 1992

National Marine Fisheries Service
2570 Dole Street
Honolulu, Hawaii

Tim Gerrodette
Christina Perrin

ADMINISTRATIVE REPORT LJ-91-36

calib.wp5
12/20/91

CALIBRATION OF SHIPBOARD ESTIMATES OF DOLPHIN SCHOOL SIZE FROM AERIAL PHOTOGRAPHS

Tim Gerrodette

Christina Perrin

Southwest Fisheries Science Center
La Jolla, CA 92038

ABSTRACT

Accurate counts of 171 dolphin schools based on high quality aerial photographs were compared with shipboard observer estimates of the same schools. Several methods of adjusting observer estimates were assessed using a cross-validatory approach, and a best procedure was determined for each observer. For 13 of 23 observers, estimates of school size were unbiased and were not improved by calibration against schools of known size; for the other 10, estimates were improved by calibration. Most of the observers whose estimates were improved by calibration tended to estimate low. A weighted mean of the observers' calibrated estimates provided the best combined estimate of school size, based on minimum squared error. Overall, dolphin school size was estimated without significant bias by observers on research vessels, although there was considerable variability in the estimates.

INTRODUCTION

Many types of animal survey require that the size of a group of animals be estimated. Obtaining accurate estimates of group size is important because the accuracy and precision of the estimates of group size directly affects the estimate of total abundance. If a population is completely censused, the sizes of the groups are added together for an estimate of total abundance. More frequently, a population is estimated by sampling. For animals that form social groups, the group is often the basic unit on which density is estimated. The total density of individuals is calculated by multiplying the estimated number of groups by mean group size.

Several studies have addressed the question of variability in estimating group size in wildlife studies (LeResche and Rausch 1974, Caughley 1974, Caughley et al. 1976, Erwin 1982, Ryan and Cooper 1989). Evaluating the accuracy of estimates of group size is a more difficult task, because it requires some method of determining true size (Jolly 1969). The method of "truthing" is, almost by definition, more difficult and/or more expensive than the estimation procedure. Many factors may affect the accuracy of estimates: difficulties associated with the visual perception of a multitude of objects, the inability to see all animals in a group because some are hidden from view, and problems with counting (not estimating) the animals that can be seen (Graham and Bell 1969).

The specific origins of this paper lie in a controversy that arose more than a decade ago. During the 1960s and 1970s, tuna seiners killed thousands of dolphins during the commercial fishing operations. Estimates of dolphin school sizes made by biologists from research vessels were considerably smaller than estimates made by fishermen from seiners, and these led to lower estimates for the number of remaining dolphins. At the time, US tuna fishermen were strongly resisting new regulations intended to reduce the mortality of dolphins. The fishermen argued that their estimates of school size were correct -- after all, they had far more experience at sea than the biologists. There was no independent source of information on school size that could resolve the difference. As a result, the SWFSC began to use a helicopter from which to take aerial photographs of the dolphin schools. Under good conditions, the dolphins in a school can be clearly seen and accurately counted from a photograph.

Marine mammal surveys were conducted annually from the NOAA research vessels David Starr Jordan and McArthur from 1986 through 1990. The primary purpose of the surveys was to monitor the abundance of dolphins killed in the tuna purse-seine fishery in the eastern tropical Pacific Ocean (Holt and Sexton 1990). The 4 dolphin species most affected by the fishery, herein referred to as target species, were spotted (Stenella attenuata), spinner (S. longirostris), striped (S. coeruleoalba), and common (Delphinus delphis) dolphins. The surveys were on a very large scale, each ship spending 4 months at sea covering a 19 million km^2 study area

in the eastern tropical Pacific Ocean (Wade and Gerrodette in press).

Over the 4 years of helicopter use (1987-90), approximately 10% of all target dolphin schools detected were photographed. The other 90% of the dolphin schools were either unphotographed or, if photographed, the quality of the photographs did not allow school size to be determined accurately enough (a few cases). The school sizes of these 90% were therefore unknown, although shipboard observer estimates were available. The problem, then, was to use the observed relation between estimated and known school sizes for the 10% of the schools that were photographed to improve the estimates for the other 90% that were not.

This paper examines and evaluates several different methods of adjusting observers' estimates of dolphin size, using the information available from aerial photographs.

METHODS

From 1987 through 1990 a Hughes 500D helicopter was carried on the David Starr Jordan. In suitable sea and sun angle conditions, the helicopter took medium-format (127 mm), motion-compensated aerial photographs of dolphin schools using a KA45A camera. In the laboratory, dolphins in the photographs were independently and repetitively counted by three different technicians using dissecting microscopes. Dolphin schools were considered to be of

known size and used in the present analysis only if the three counts agreed closely according to procedures in Gilpatrick et al. (1991). The mean of the three best counts was taken as the true school size.

The analysis proceeded in two stages. First, several different procedures for calibrating each observer's estimates were examined. Second, different methods of combining the individual observer estimates into a mean estimate of school size were considered. The performances of the various methods were assessed using computer-intensive resampling techniques. Linear regression was the basic method of analysis employed. In standard regression models, the predictor variable (X) is assumed to be measured without error. However, school size, as the mean of three counts of an aerial photograph, did have measurement error. Therefore, both standard models and models that included error in the X variable were considered (Fuller 1987). Both school size and observer estimate were log transformed to achieve normality and homoscedasticity.

Selection for off-duty and non-target sightings

The set of photographed dolphin schools differed by two factors that were not present in the set of unphotographed dolphin schools to which the calibrations were to be applied: (1) the photographed schools contained both on- and off-duty estimates, and (2) the photographed schools contained both target and non-target

species. For each observer, therefore, we tested whether the observer estimated school size differently while on or off duty, or estimated differently between target and non-target species. If significant differences were indicated at the α=0.05 level, school size estimates made while off duty and/or on non-target species were not used in calibrating that observer.

Following Neter et al. (1990), tests for differences between on- and off-duty sightings were based on the differences in residual sum of squares using the full model

$$\log B = \beta_0+\beta_1 \log X_1+\beta_2 X_2+\beta_3 X_1 X_2+\sum_{j=4}^{p}\beta_j X_j + \epsilon$$

versus the reduced model

$$\log B = \beta_0+\beta_1 \log X_1+\sum_{j=4}^{p}\beta_j X_j + \epsilon$$

where B = observer best estimate

X_1 = true school size

X_2 = 1 if on duty, 0 if off duty

X_j = other categorical (0 or 1) factors for species, sea state, and year, with all first-order interaction terms with X_1.

A similar model tested for differences between target and nontarget species, except that the reduced model did not contain terms for target/non-target categories, but did contain terms for on and off

duty sightings. Tests for equality of variance between the groups were also carried out.

Selection of weights for estimates

For each dolphin school, each observer made three estimates: best (B), high (H), and low (L). A linear combination of the three estimates was selected as the observer's weighted estimate of school size for school i:

$$\tilde{S}_i = w_1 B_i + w_2 H_i + w_3 L_i, \qquad \sum_{j=1}^{3} w_j = 1.$$

Optimal weights w_j were computed by minimizing the residual sum of squares in the regression of log $\tilde{S}$ against log X, where X = true school size. Weights were computed separately for each observer using the downhill simplex method of Nelder and Mead (Press et al., 1989). To ensure that a global minimum was located, 20 searches were conducted, each time starting the corners of the simplex at different random points of the parameter space.

Individual observer calibration

For each observer, three different calibration procedures to estimate school size $\hat{S}$ were considered.

Procedure 1:

$$\hat{S}_{1,i} = B_i = \text{best estimate of school } i.$$

The first procedure simply used the observer's unadjusted best estimate of each school. If no aerial photographs had been available, this estimate would have been used. Thus, procedure 1 provided a null case against which to measure any improvement obtained by calibration of estimates with schools of known size in procedures 2 and 3.

Procedure 2:

$$\hat{S}_{2,i} = \text{antilog}\left[\frac{1}{b_1}(\log \tilde{S}_i - b_0)\right],$$

where the coefficients b were computed from the regression

$$\log \tilde{S}_i = b_0 + b_1 \log X_i.$$

The second procedure used coefficients from a bivariate regression of weighted estimates of school size ($\tilde{S}$) against true school sizes (X) to "correct" the observer's estimates.

Procedure 3:

$$\hat{S}_{3,i} = \text{antilog}\left[\frac{1}{b_1}(\log \tilde{S}_i - b_0 - \sum_{j=2}^{5} b_j X_j)\right],$$

where the coefficients b were computed from the regression

$$\log \tilde{S}_i = b_0 + b_1 \log X_{1,i} + \sum_{j=2}^{5} b_j X_{ji},$$

where $X_{1,i}$ = true size of school i,

$X_{2,i}$ = 1 if sea state ≤ Beaufort 2 for school i, 0 otherwise,

$X_{3,i}$ = 1 if school i was seen in 1987, 0 otherwise,

$X_{4,i}$ = 1 if school i was seen in 1988, 0 otherwise,

$X_{5,i}$ = 1 if school i was seen in 1989, 0 otherwise.

The third procedure was thus similar to the second, but used coefficients from a multivariate regression of weighted estimates of school size ($\tilde{S}$) against true school sizes (X_1), with sea state and years as covariates, to "correct" the observer's estimates. The covariates allowed for possible differences in the way an observer estimated dolphin school size in different sea conditions or in different years. Correction factors for bias resulting from logarithmic regression (Beauchamp and Olson 1973) were not applied in this case because the antilogarithm was being taken on the predictor axis.

These three procedures were assessed for each observer using a data resampling method similar to cross-validation (Stone 1974, Geisser 1975). A data point (dolphin school) was deleted from the set of known schools for that observer, and the size of the deleted school was predicted based on the remaining data points and the observer's best, high, and low estimates for that school. The prediction error was the difference between the estimate and the

true size of the deleted school. This differed slightly from the usual cross-validatory assessment (Bailey et al. 1989) by measuring error as the difference between the estimate and a known true value rather than between the estimate and an observed value. The error rate associated with estimating the size of an unknown dolphin school, given the observer's shipboard estimates, was estimated by the average square prediction error

$$ASPE_j = \frac{1}{n} \sum_{i=1}^{n} (\log X_i - \log \hat{S}_{j(i)})^2,$$

for j=1,2,3, where n is the number of schools, X_i is the true size of the i-th school, and $\hat{S}_{j(i)}$ indicates an estimate using the j-th procedure made with the i-th data point removed. This statistic thus combined components of both bias and variance. The procedure with the smallest ASPE was selected as the best procedure for each observer.

Combining individual estimates

Because observers operated in teams of three, each dolphin school had three sets of best, high, and low estimates, one set from each observer. A strong attempt was made to maintain independence and consistency of observer estimates, both within a cruise and among years. The observers did not discuss their estimates of school size with each other or with the photographic technicians in the helicopter; they were also not informed of the

results of the aerial photography.

Three methods of combining the individual observer estimates to obtain a mean estimate for the dolphin school were considered: (1) the mean of the observers' unadjusted best estimates (as before, this provided a null case); (2) the unweighted mean of the observers' calibrated estimates, using the best calibration procedure for each individual; and (3) the weighted mean of the observers' calibrated estimates, using the inverse of the ASPE for each observer as weights. The performances of these three methods were again compared by resampling, using the statistic

$$ASPE_k = \frac{1}{n} \sum_{i=1}^{n} (\log X_i - \log \bar{S}_{k(i)})^2,$$

where $\bar{S}_{k(i)}$ indicates mean school size using the k-th method above (k=1,2,3) made with the i-th data point removed.

RESULTS

A total of 171 photographed dolphin schools met the criteria for precision and quality, distributed among the four years as follows: 1987: 46, 1988: 49, 1989: 39, 1990: 37. Most of these schools had 2 sets of school size estimates, because when a school was successfully photographed, the "off-duty" team of 3 observers also made a set of estimates. Thus there was a total of 312 sets of estimates for which the true school size was known. There were

23 different individual observers. Because observers worked variable numbers of years, the number of known schools against which to calibrate each observer varied from 7 to 80, with a median of 35 (Fig. 1).

Standard regression models assuming no error in the predictor variable were sufficient for this calibration problem. The reliability ratio, the factor by which the slope is underestimated if one assumes no measurement error in X when there actually is, is $\kappa=\sigma_X/(\sigma_X+\sigma_m)$, where σ_X is the standard deviation of X and σ_m is the standard deviation of measurement error in X (Fuller 1987). Good estimates of σ_m were available from replicate counts of the same photograph. From these, the reliability ratio was computed to be $\kappa=0.989$. Thus, the attenuation was only about 1%, and school size was effectively measured without error.

Least square regressions of observer estimates against school size were computed for each observer (Fig. 1). Estimates were stratified by duty status (on or off), sea state (Beaufort ≤ 2 or ≥ 3), species (target or non-target), and year (1-4) (Fig. 1). Generally speaking, none of these factors had a consistent effect over all observers in the estimation of dolphin school size. Tests for differences in duty status and species showed that observers 9, 16, and 17 estimated differently while off duty, and observer 4 estimated non-target species differently (Table 1). The number of schools on which calibration factors were computed for these observers was reduced accordingly. All other observers showed no significant differences for these factors, and all schools were

used for their calibrations.

There were distinct individual differences among the 23 observers in both the accuracy and precision of estimating dolphin school size (Fig. 1, Table 2). The ASPEs of the unadjusted best estimates (procedure 1) ranged from 0.0193 (observer 14) to 0.2306 (observer 23), with a median of 0.0556. For 13 of the 23 observers, the unadjusted best estimate was superior to either of the calibration methods attempted (Table 2). Of the remaining 10 observers, 5 of them were best calibrated using simple bivariate regression (procedure 2) and 5 using regression with sea state and year covariates (procedure 3) (Table 2). For some observers, the improvement in estimating school size using calibrated estimates was substantial (observers 9 and 23, for example). After selection of the best procedure for each observer (marked with asterisks in Table 2), the ASPEs ranged from 0.0193 to 0.0929, with a median of 0.0547.

The estimates of dolphin school size for each observer, using the best procedure for each observer, are plotted in Fig. 2. Most of the 10 observers who benefitted from a calibration procedure tended to estimate low (Fig. 2, open symbols), and the calibration procedure raised most estimates. Observer 16 was an exception, with most estimates lowered as a result of the calibration procedure. After selection of the best procedure for each observer, estimates of school size were generally near the 1:1 line (Fig. 2, filled symbols).

Error rates for mean school size estimates (Table 3) were

generally lower than the error rates for the individual observer estimates (Table 2). The weighted mean of the calibrated estimates had a lower ASPE than the other methods of computing mean school size (Table 3). There was a reduction in the error rate from 0.0433 for the mean of the unadjusted best estimates to 0.0376 for the weighted mean of the calibrated estimates. Thus, use of the aerial photographs resulted in an overall improvement of about 13% in the squared error rate. The weighted mean calibrated estimates appeared randomly distributed about the known true school sizes (Fig. 3).

DISCUSSION

Rapid estimation of the number of objects in a cluster or group is a difficult visual task. In wildlife studies, the task is frequently made more difficult because the groups are composed of moving animals. Furthermore, there is often only a limited time that the animals can be seen. For the dolphins considered in this study, there is the additional difficulty that not all animals in the school are visible at one time. The dolphins are visible only when they surface to breathe, and they do not all breathe at the same time.

However, the size of dolphin schools can be assessed from aerial photographs. Previous studies (Clark 1984, Scott et al. 1985) have shown that counts of dolphin schools from photographs are precise; they also agree with counts of dolphin schools made

in the backdown channel of the purse seine (Allen et al. 1980), so it is also thought that photographs are accurate. The precision and consistency of the dolphin photo counting procedure for the dolphin schools used in this study has been evaluated by Gilpatrick et al. (1991). In general, replicate counts of photographs are highly consistent.

However, estimates of dolphin school size, whether from shipboard or aerial, are considerably more variable. Clark (1984) and Scott et al. (1985) conducted earlier analyses of visual and photographic estimates of dolphin school size. Photographs were taken from helicopters and airplanes with several kinds of cameras. Estimates of school size were made from the air and aboard ship, and counts were made at the backdown channel. The general conclusions of these previous studies were that counts from aerial photographs were precise and that the relative error of visual estimates of dolphin school size differed significantly among observers. Studies on other species (LeResche and Rausch 1974, Caughley et al. 1976, Erwin 1982) have also shown substantial individual variation among observers. Clark (1984) and Scott et al. (1985) concluded that it was not possible to calibrate the "population" of observers because of individual differences, but that data were insufficient to derive individual calibration factors.

In this study, the 23 observers did remarkably well at estimating dolphin school size under actual survey conditions. Over half (13/23) did so well that their estimates could not be

improved by calibration against known schools. Using a regression equation for calibration actually made their estimates worse. This may seem like a paradox, because if the best estimates are already very good the regression line should cause no adjustment, and estimates based on it should be at least as good as the unadjusted estimates. However, this is not so. It is the accuracy of estimating a future unknown point that is of interest. If the statistical model is overspecified, it will be too heavily influenced by the peculiarities of the sample. The cross-validatory assessment of different predictors can estimate the penalty associated with overspecification (Stone 1974).

Clark (1984) and Scott et al. (1985) found that small dolphin schools were generally estimated accurately, but that large schools tended to be underestimated. The tendency to underestimate the size of a group of objects is apparently common (Erwin 1982). In this study, most of the 10 observers whose estimates could be improved by calibration tended to underestimate school size. With the exception of observer 16, the calibration procedure generally had the effect of raising estimates closer to true school size (Fig. 2). For some of these 10 observers, the improvement was slight, but for others the improvement was dramatic. For example, calibration of observer 23 reduced the estimated squared error rate from 0.2306 to 0.0796. Because these error rates apply to a logarithmic parameter space, comparison of calibrated and uncalibrated mean school sizes with true mean school size is easier to appreciate. For the 12 schools seen by observer 23, the true

mean size of the 12 schools was 234.1 dolphins. The mean of this observer's 12 unadjusted "best" estimates was 70.5, a considerable underestimate. After calibration, the mean of the 12 calibrated estimates was 260.3, not perfect, but much closer to the true value.

However, even after calibration, there still remains a considerable amount of error in any single estimate. A typical value for the error rate for the calibrated estimates was about 0.05 (Table 2). Translated into an arithmetic scale, this means that the standard error of a single estimate is approximately 1.7 times the estimate. If logarithms of estimates are normally distributed, the average observer's estimate of a school of 100 dolphins, after calibration, will fall between 35 and 287 dolphins with probability 0.95.

The error rate was reduced by taking the mean of 3 estimates, an expected result if estimation errors by observers are random and independent. The error rate of the mean of the calibrated estimates was less than the error rate of the unadjusted best estimates (Table 3). Use of a weighted mean resulted in a slight further improvement over the unweighted case because estimates by "better" observers were given more weight.

REFERENCES CITED

Allen, R. L., D. A. Bratten, J. L. Laake, J. F. Lambert, W. L. Perryman, and M. D. Scott. 1980. Report on estimating the size of dolphin schools, based on data obtained during a charter cruise of the M/V Gina Anne, October 11- November 25, 1979. Inter-Amer. Trop. Tuna Commn. Data Report 6, 28 p.

Bailey, R. A., S. A. Harding, and G. L. Smith. 1989. Cross-validation. Encyclopedia of Statistical Science, Supp. Vol.

Beauchamp, John J., and Jerry S. Olson. 1973. Corrections for bias in regression estimates after logarithmic regression. Ecology 54: 1403-1407.

Caughley, G. 1974. Bias in aerial survey. J. Wildl. Manage. 38: 921-933.

Caughley, G., R. Sinclair, and D. Scott-Kemmis. 1976. Experiments in aerial survey. J. Wildl. Manage. 40: 290-300.

Clark, William G. 1984. Analysis of variance of photographic and visual estimates of dolphin school size. SWFSC Admin. Rpt. LJ-84-11C, 36 p.

Erwin, R. M. 1982. Observer variability in estimating numbers: an experiment. J. Field. Ornithol. 53: 159-167.

Fuller, Wayne A. 1987. Measurement Error Models. John Wiley & Sons, New York.

Geisser, S. 1975. The predictive sample reuse method with applications. J. Amer. Stat. Assn. 70: 320-328.

Gilpatrick, James. W., Jr., Morgan S. Lynn, and Robin L. Westlake.

1991. Image interpretation and reader variability in dolphin school size estimates made from aerial photographs. MS.

Graham, A., and R. Bell. 1969. Factors influencing the countability of animals. E. Afr. Agric. For. J. 34: 38-43.

Holt, R. S., and S. N. Sexton. 1990. Monitoring trends in dolphin abundance in the eastern tropical Pacific using research vessels over a long sampling period: Analyses of 1986 data, the first year. Fish. Bull. 88: 105-111.

Jolly, G. M. 1969. The treatment of errors in aerial counts of wildlife populations. E. Afr. Agric. For. J. 34: 50-55.

LeResche, R., and R. Rausch. 1974. Accuracy and precision of aerial moose censusing. J. Wildl. Manage. 38: 175-182.

Neter, John, William Wasserman, and Michael H. Kutner. 1990. Applied Linear Statistical Models. Irwin, Homewood, IL.

Press, William H., Brian P. Flannery, Saul A. Teukolsky, and William T. Vetterling. 1989. Numerical Recipes. Cambridge University Press, Cambridge.

Ryan, P. G., and J. Cooper. 1989. Observer precision and bird conspicuousness during counts of birds at sea. S. Afr. J. Mar. Sci. 8: 271-276.

Scott, Michael D., Wayne L. Perryman, and William G. Clark. 1985. The use of aerial photographs for estimating school sizes of cetaceans. Bull. Inter-Amer. Trop. Tuna Commn. 18: 383-404.

Stone, M. 1974. Cross-validatory choice and assessment of statistical predictions. J. Roy. Stat. Soc., Ser. B, 36: 111-133.

Wade, P., and T. Gerrodette. In press. Monitoring trends in dolphin abundance in the eastern tropical Pacific: analysis of five years of data. Rep. Int. Whal. Comm.

Table 1. Statistics of tests for differences between on- and off-duty and target and non-target estimates for each of 23 observers. SS=residual sum of squares, df=degrees of freedom, MSE=mean square error, F=test statistic. Results significant at the α=0.05 level are marked with an asterisk (*).

	Full model			Reduced model					
				On/off duty			Target/nontarget sp.		
Obs.	SS	df	MSE	SS	df	F	SS	df	F
1	1.1030	23	0.0480	1.2173	25	1.19	1.3128	24	2.19
2	2.3849	35	0.0681	2.4306	37	0.34	2.5578	36	1.27
3	1.4697	29	0.0507	1.4947	31	0.25	1.4787	30	0.09
4	1.3790	24	0.0575	1.3853	26	0.06	1.8474	25	4.07*
5	1.6959	37	0.0458	1.9015	39	2.25	1.7897	39	1.02
6	1.5772	45	0.0350	1.7205	47	2.05	1.6988	47	1.74
7	2.5000	59	0.0424	2.6608	61	1.90	2.5696	61	0.82
8	1.2628	25	0.0505	1.3868	27	1.23	1.5055	27	2.40
9	0.2316	14	0.0178	0.4833	15	7.07*	0.2833	14	1.45
10	2.2732	63	0.0361	2.3654	65	1.28	2.4269	65	2.13
11	2.9060	59	0.0493	2.9972	61	0.92	3.0386	61	1.34
12	3.2885	49	0.0671	3.3876	51	0.74	3.4274	51	1.04
13	0.3787	6	0.0631	0.3801	8	0.01	0.4022	7	0.19
14	0.0026	1	0.0026	0.0778	3	14.46	0.0027	2	0.02
15	0.4290	10	0.0429	0.4714	12	0.49	0.4290	10	0.00
16	0.6638	25	0.0266	0.9376	27	5.15*	0.6641	27	0.01
17	3.4412	66	0.0521	3.8589	68	4.01*	3.5676	68	1.21
18	1.0937	37	0.0296	1.1012	39	0.13	1.4510	38	0.36
19	2.0084	48	0.0418	2.0571	50	0.58	2.0238	50	0.18
20	0.6697	18	0.0372	0.7921	20	1.65	0.6973	20	0.37
21	1.3147	24	0.0548	1.3223	26	0.07	1.4936	25	1.63
22	0.8744	17	0.0514	0.8904	19	0.16	0.9291	18	0.53
23	0.1108	6	0.0185	0.1411	8	0.82			

Table 2. Evaluation of 3 different calibration procedures (Proc.) for each of 23 observers. N=no. of calibration schools, w=weights for best, high, and low estimates, b=regression coefficients (see text), ASPE=average square prediction error. Asterisk (*) marks best procedure for each observer, based on lowest ASPE.

Obs.	N	Proc.	w1	w2	w3	b0	b1	b2	b3	b4	b5	ASPE
1	32	1*	1.00	.00	.00							.0556
		2	.37	.00	.63	-.018	.986					.0692
		3	.37	.00	.63	-.040	.911	.172	.000	.000	.184	.0749
2	44	1	1.00	.00	.00							.0887
		2	.00	.11	.89	-.094	.963					.0829
		3*	.00	.11	.89	.037	.927	.080	.000	.000	-.168	.0824
3	38	1	1.00	.00	.00							.1289
		2*	.00	.58	.42	.059	.846					.0645
		3	.00	.58	.42	.082	.836	.032	.000	.000	-.021	.0724
4	29	1*	1.00	.00	.00							.0663
		2	.00	1.00	.00	.886	.680					.1063
		3	.00	1.00	.00	.818	.694	.060	.000	.000	.000	.1094
5	49	1*	1.00	.00	.00							.0498
		2	.00	.00	1.00	.154	.902					.0578
		3	.00	.00	1.00	-.013	.930	.028	.117	.167	.000	.0569
6	55	1*	1.00	.00	.00							.0668
		2	.76	.00	.24	.285	.823					.0918
		3	.75	.00	.25	.536	.763	-.074	-.188	.000	.000	.1005
7	73	1*	1.00	.00	.00							.0469
		2	.00	.88	.12	.175	.908					.0530
		3	.00	.88	.12	.136	.913	.025	.023	.007	.032	.0578
8	35	1*	1.00	.00	.00							.0512
		2	.00	.00	1.00	.055	.952					.0586
		3	.00	.00	1.00	-.062	.988	.107	.003	.000	.000	.0577
9	10	1	1.00	.00	.00							.2000
		2*	.00	1.00	.00	.319	.648					.0679
		3	.00	1.00	.00	.314	.649	.005	.000	.000	.000	.1115
10	75	1*	1.00	.00	.00							.0556
		2	.00	.76	.24	.513	.816					.0593
		3	.00	.76	.24	.629	.796	.069	-.207	-.078	-.090	.0603
11	73	1*	1.00	.00	.00							.0547
		2	.37	.63	.00	.308	.836					.0723
		3	.37	.63	.00	.226	.854	.039	.109	.025	-.007	.0737

Table 2 (cont'd)

Obs.	N	Proc.	w1	w2	w3	b0	b1	b2	b3	b4	b5	ASPE
12	61	1	1.00	.00	.00							.0800
		2*	.00	.00	1.00	-.210	.995					.0642
		3	.00	.00	1.00	-.062	.966	.018	-.188	-.065	.000	.0718
13	13	1	1.00	.00	.00							.1956
		2*	.00	1.00	.00	.519	.688					.0676
		3	.00	1.00	.00	.365	.718	.112	.000	.000	.000	.0848
14	7	1*	1.00	.00	.00							.0193
		2	1.00	.00	.00	.133	.976					.0291
		3	1.00	.00	.00	.148	.980	-.038	.000	.000	.000	.0545
15	16	1*	1.00	.00	.00							.0336
		2	1.00	.00	.00	-.069	1.044					.0423
		3	1.00	.00	.00	.118	.988	-.118	.000	.000	.000	.0482
16	21	1	1.00	.00	.00							.0386
		2	.17	.00	.83	-.047	1.062					.0304
		3*	.17	.00	.83	-.236	1.105	.158	.071	.000	.000	.0242
17	38	1	1.00	.00	.00							.0561
		2	.00	.00	1.00	-.340	1.073					.0362
		3*	.00	.00	1.00	-.374	1.113	.118	-.243	-.148	-.005	.0317
18	46	1	1.00	.00	.00							.0994
		2	.00	.00	1.00	-.049	.847					.0475
		3*	.00	.00	1.00	.013	.836	.073	.000	.000	-.138	.0413
19	60	1*	1.00	.00	.00							.0377
		2	.58	.00	.42	-.044	1.001					.0417
		3	.58	.00	.42	.000	.998	.060	-.105	-.079	.000	.0459
20	28	1*	1.00	.00	.00							.0494
		2	1.00	.00	.00	.149	.874					.0629
		3	1.00	.00	.00	.014	.880	-.001	.148	.000	.000	.0639
21	33	1	1.00	.00	.00							.0963
		2	.00	.72	.28	.137	.879					.0964
		3*	.00	.72	.28	.112	.828	.041	.000	.000	.302	.0928
22	23	1*	1.00	.00	.00							.0478
		2	.00	.00	1.00	.079	.929					.0564
		3	.00	.00	1.00	.018	.915	.125	.000	.000	.000	.0576
23	12	1	1.00	.00	.00							.2306
		2*	.00	.00	1.00	.514	.563					.0796
		3	.00	.00	1.00	.530	.566	-.055	.000	.000	.000	.0856

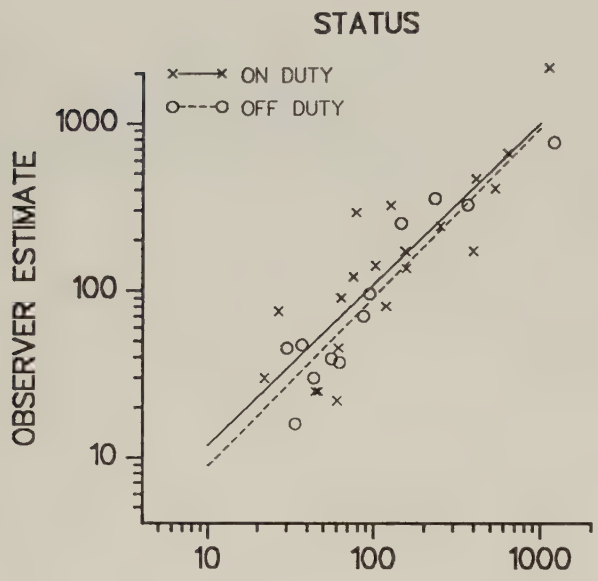
Table 3. Average squared prediction error (ASPE) for 3 different methods of computing mean dolphin school size from individual observer estimates. Calibrated estimates refer to estimates using the best calibration procedure for each observer. The weighted mean used the inverse of each observer's ASPE from Table 2 as a weighting factor.

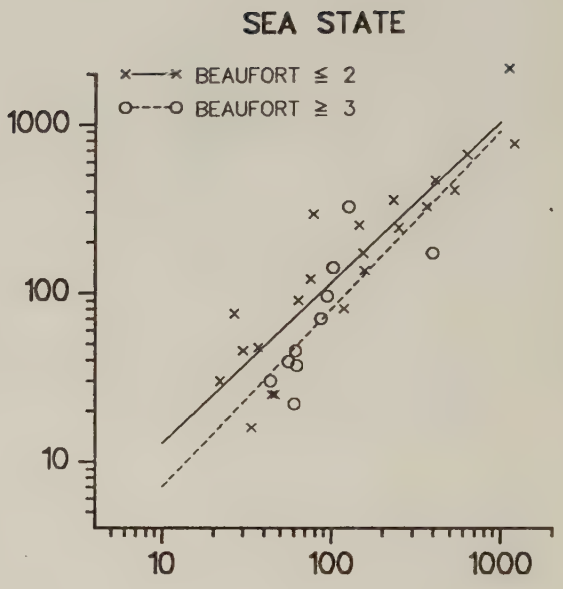
Method of computing mean	ASPE
Mean of best estimates	0.04327
Mean of calibrated estimates	0.03905
Weighted mean of calibrated estimates	0.03756

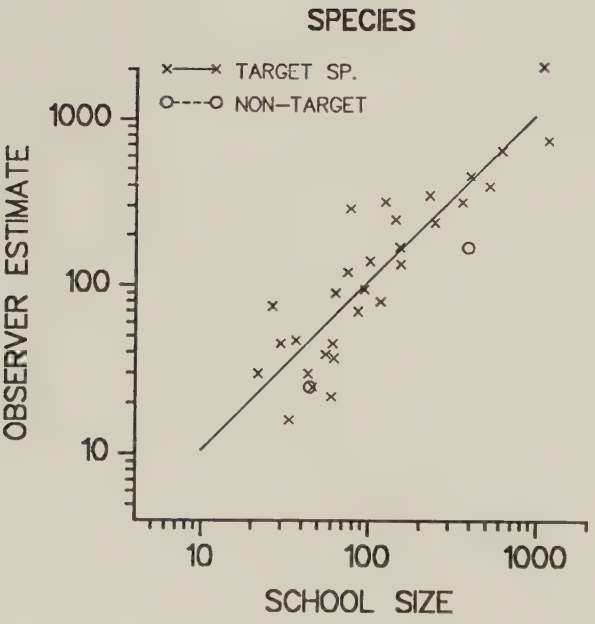
Fig. 1: Observer estimates of dolphin school size plotted against known school size for the 23 observers in this study. For each observer, best estimates are stratified by 4 factors: duty status, sea state, species of cetacean, and year of estimate. Least square regression lines are plotted.

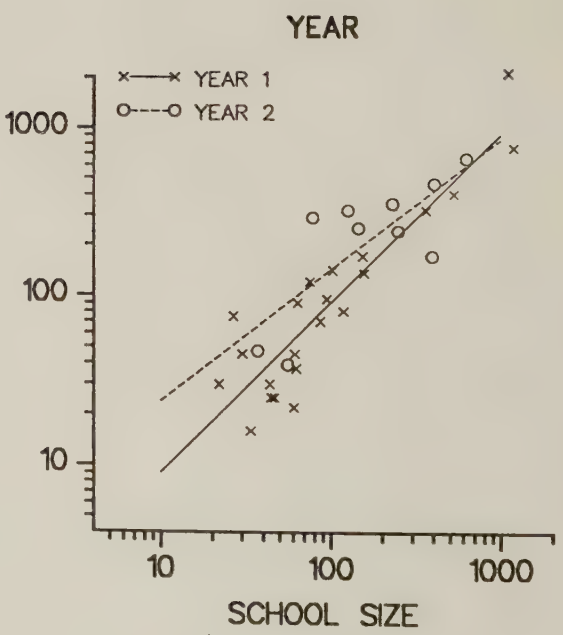
OBSERVER 1

N = 32



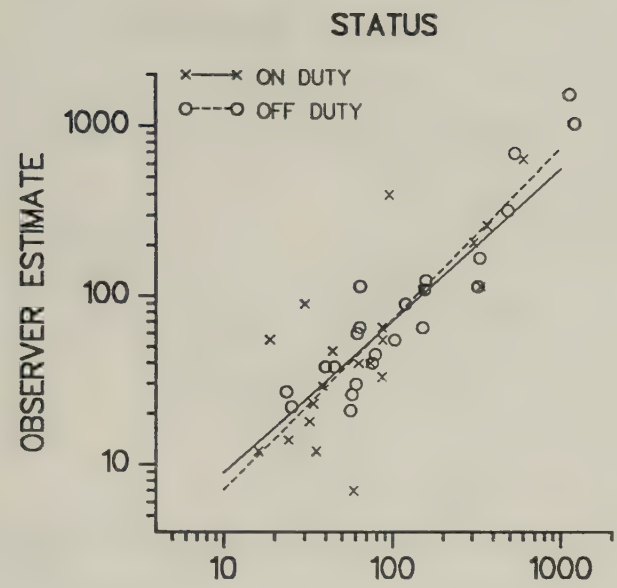


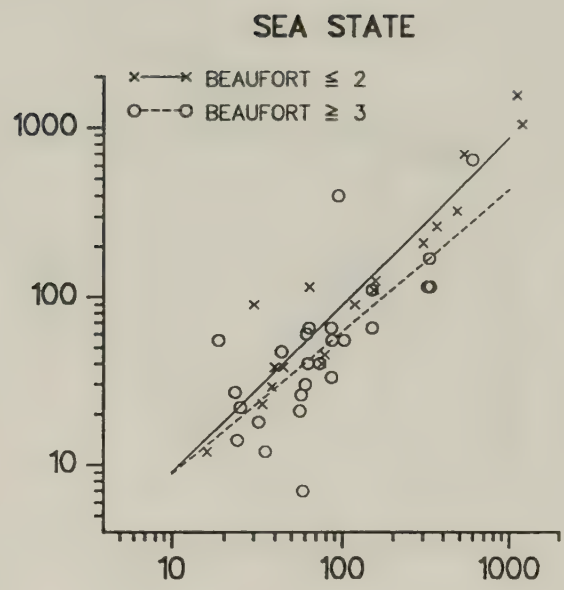


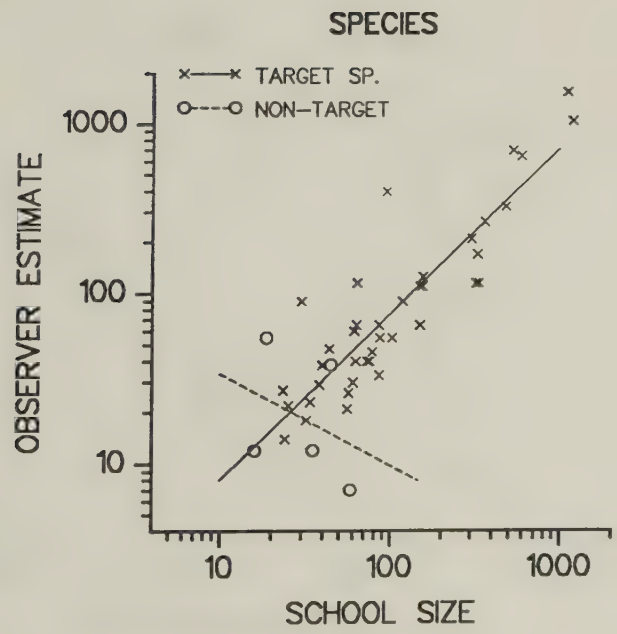


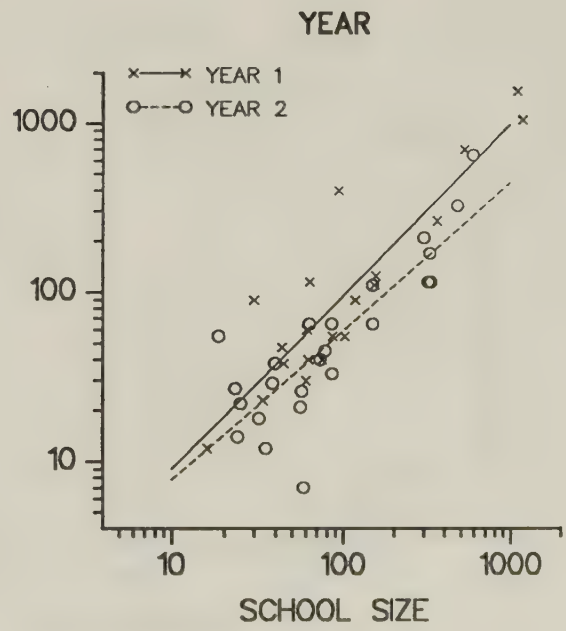
OBSERVER 2

N = 44



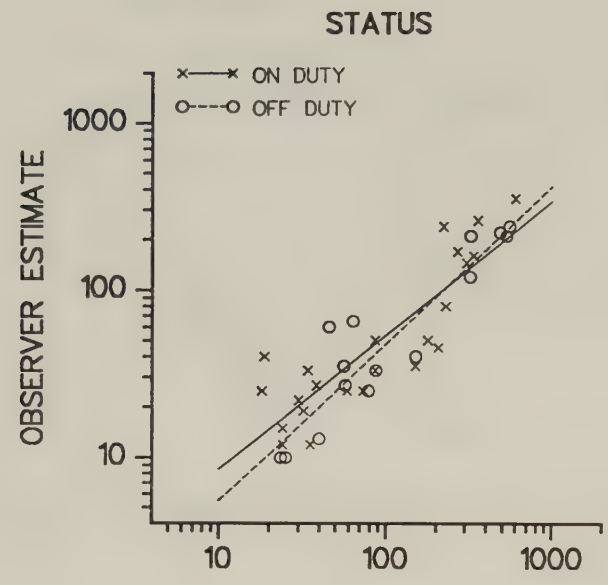


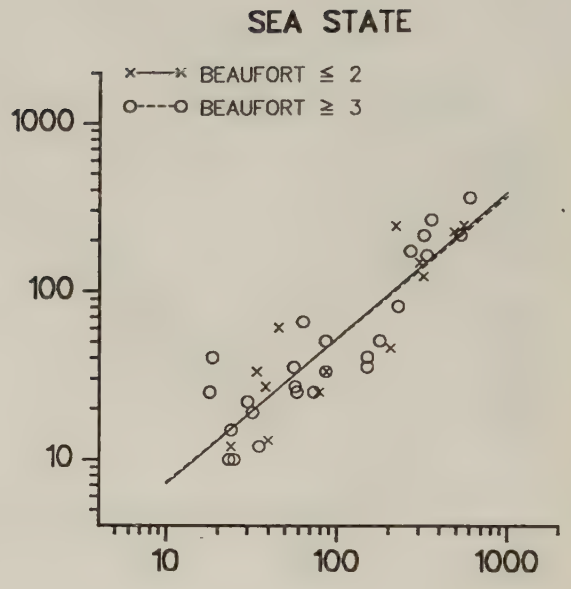


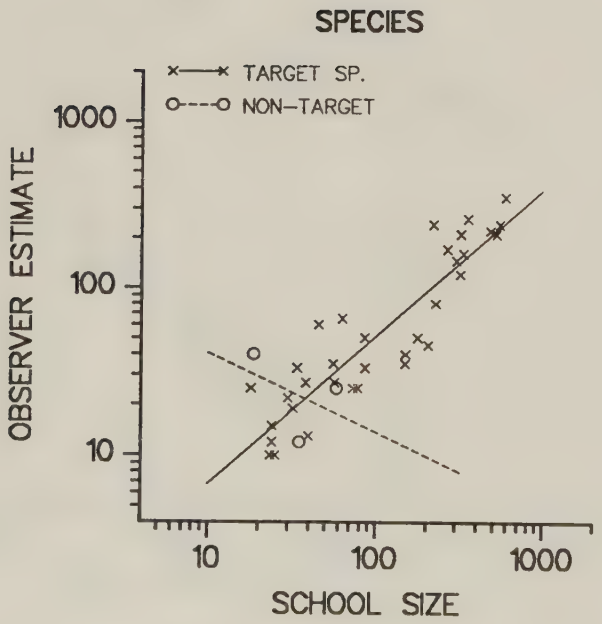


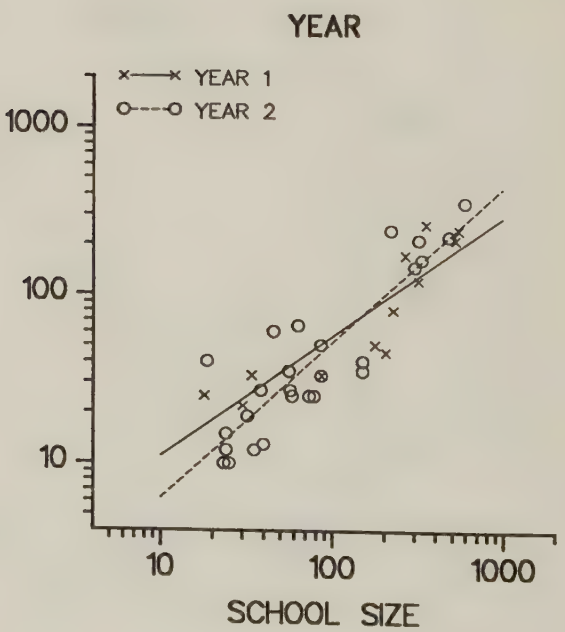
OBSERVER 3

N = 38



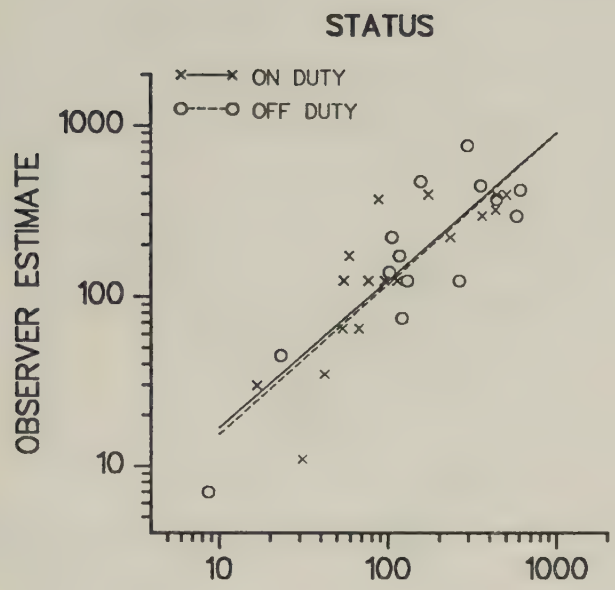


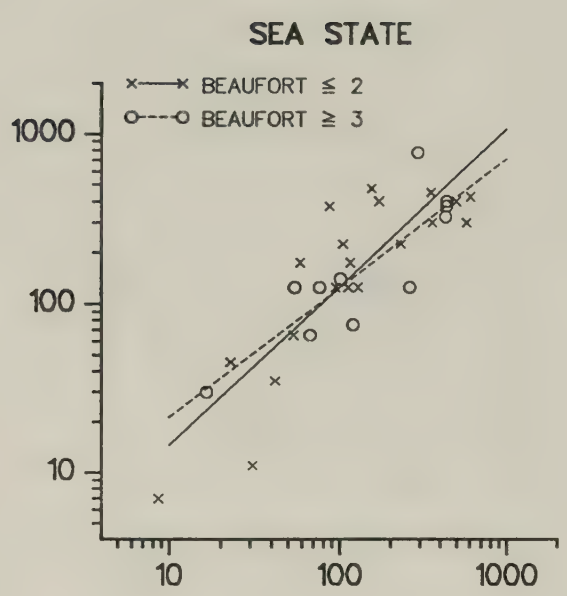


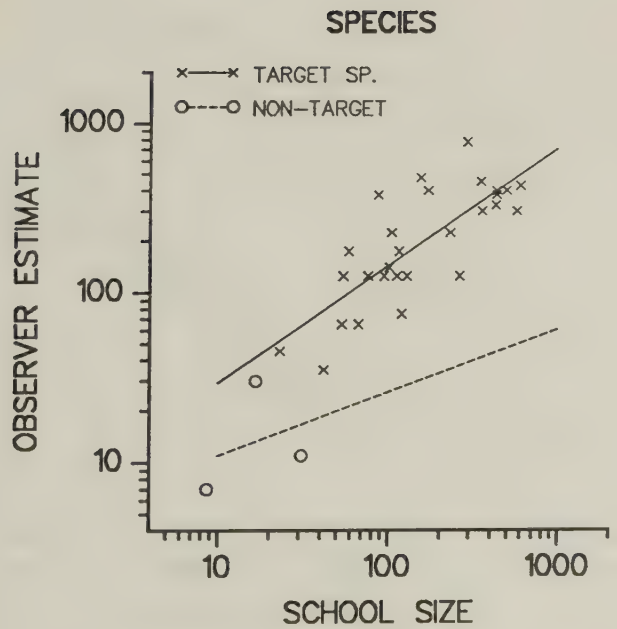


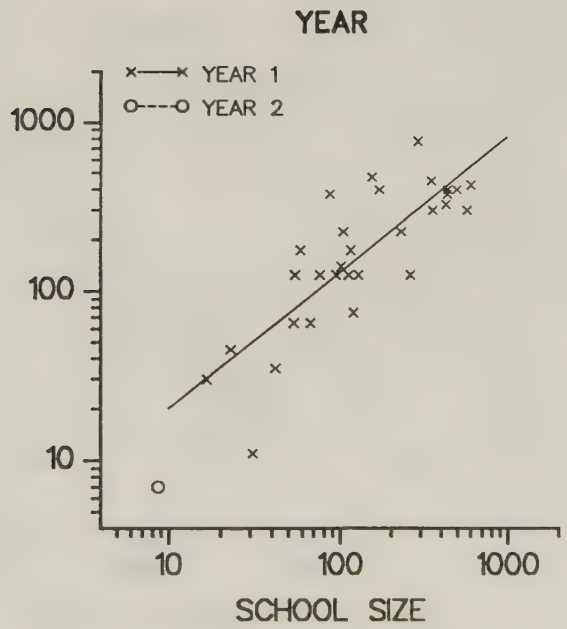
OBSERVER 4

N = 31



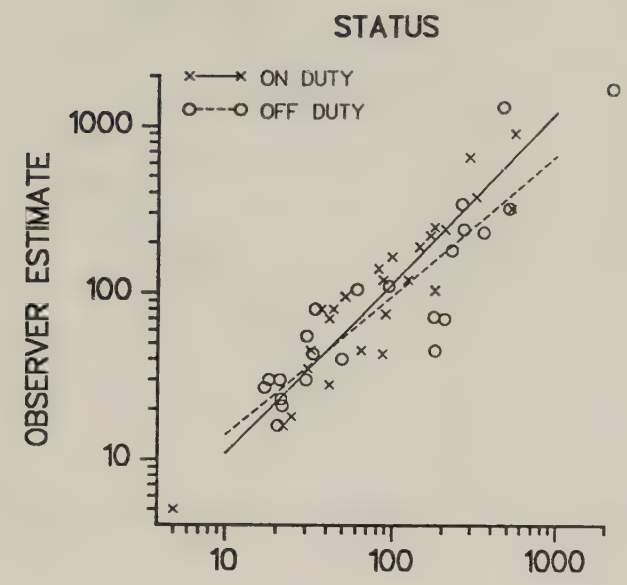


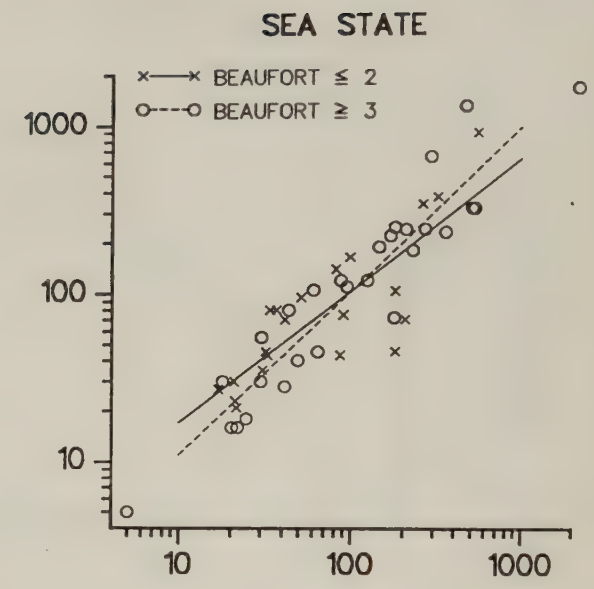


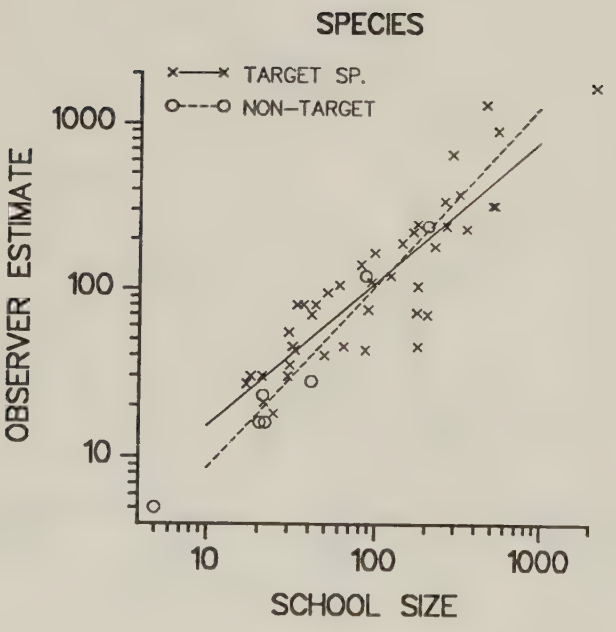


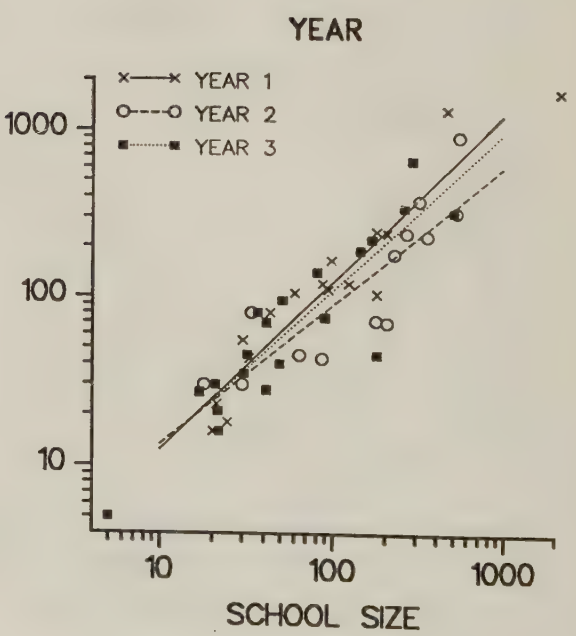
OBSERVER 5

N = 49



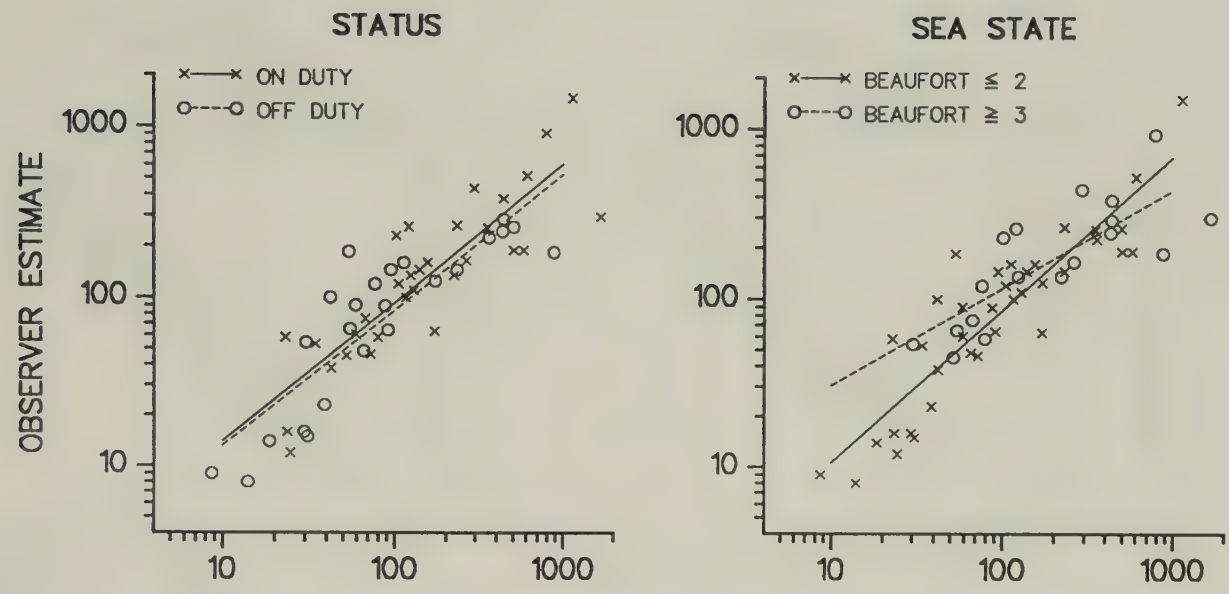


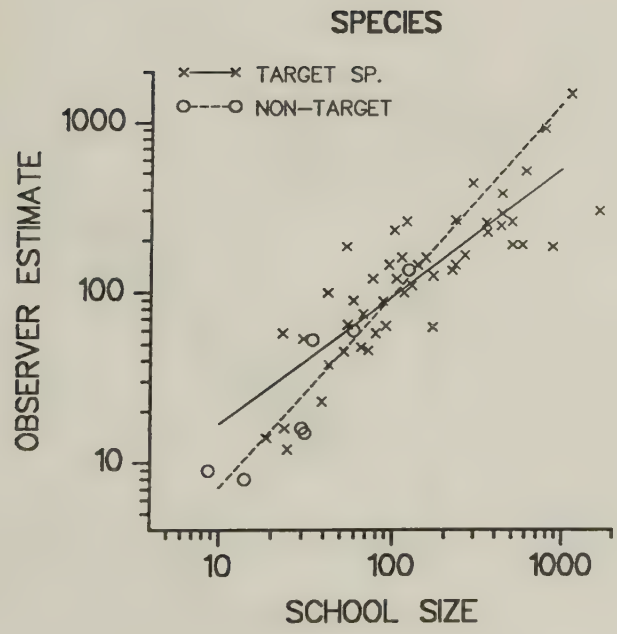


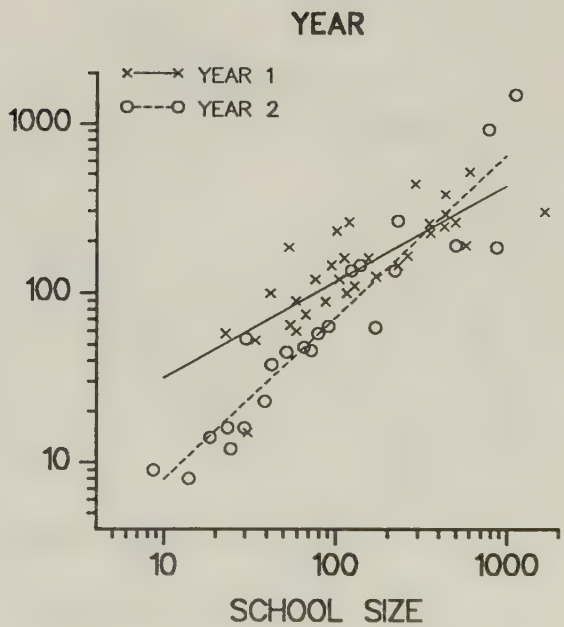


OBSERVER 6

N = 55

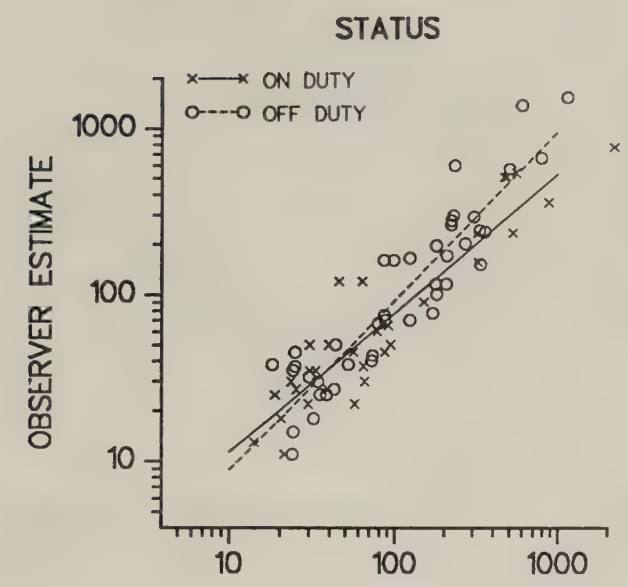


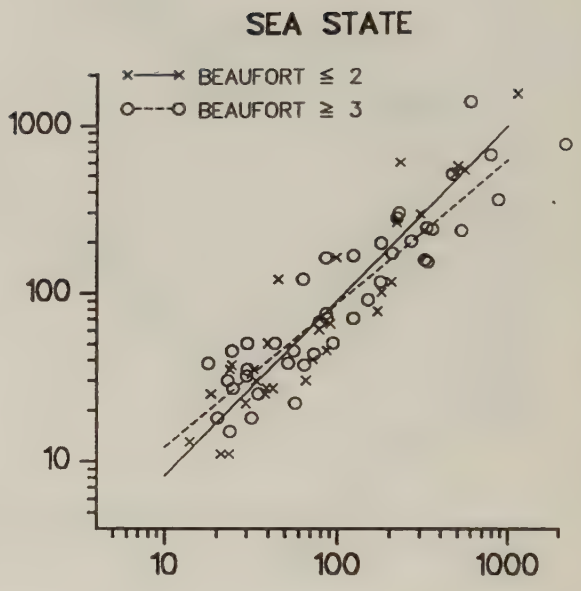


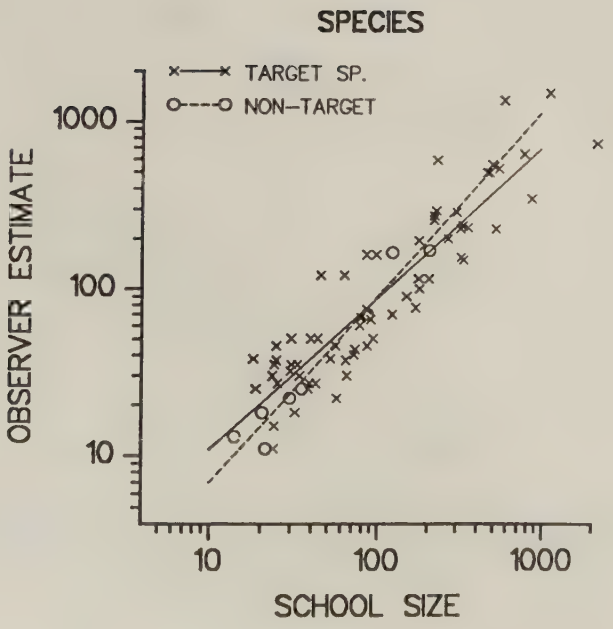


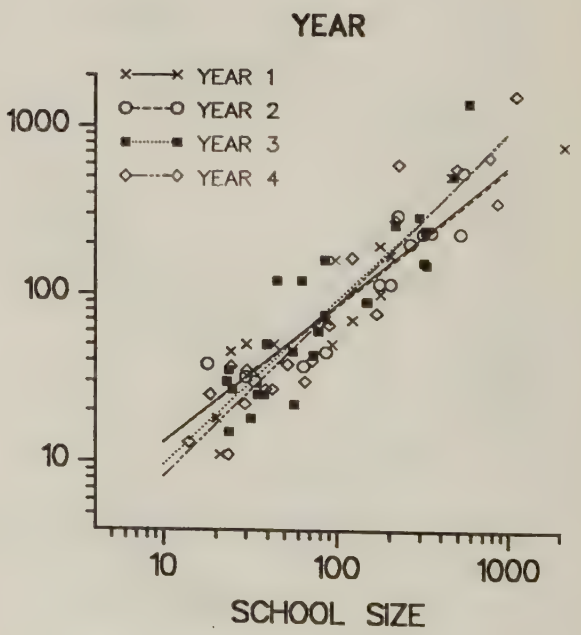
OBSERVER 7

N = 73



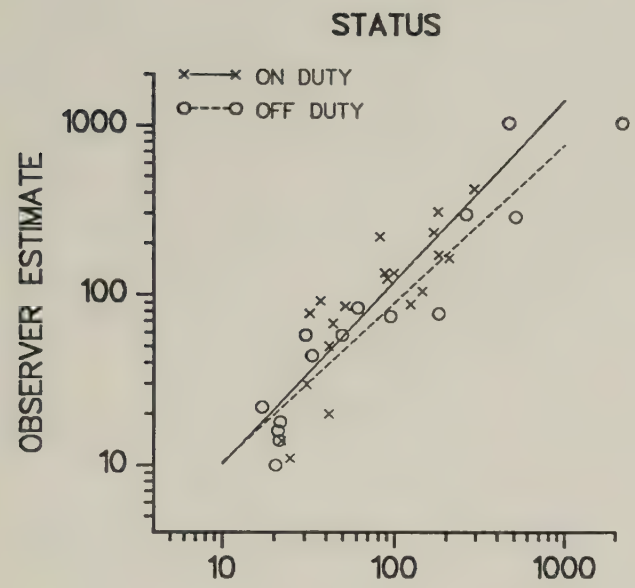


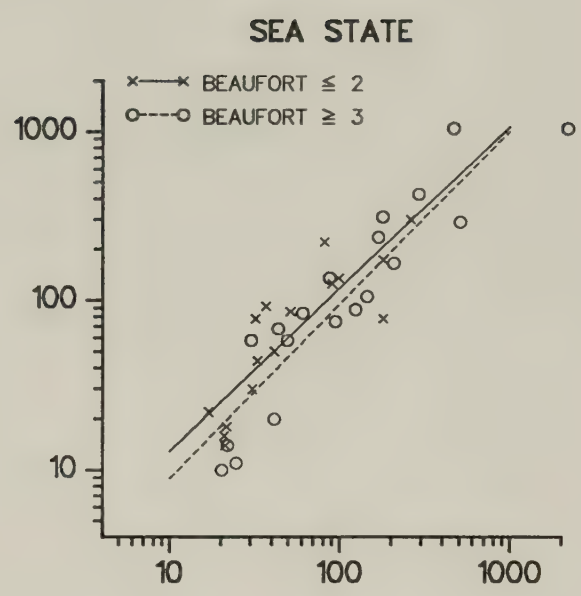


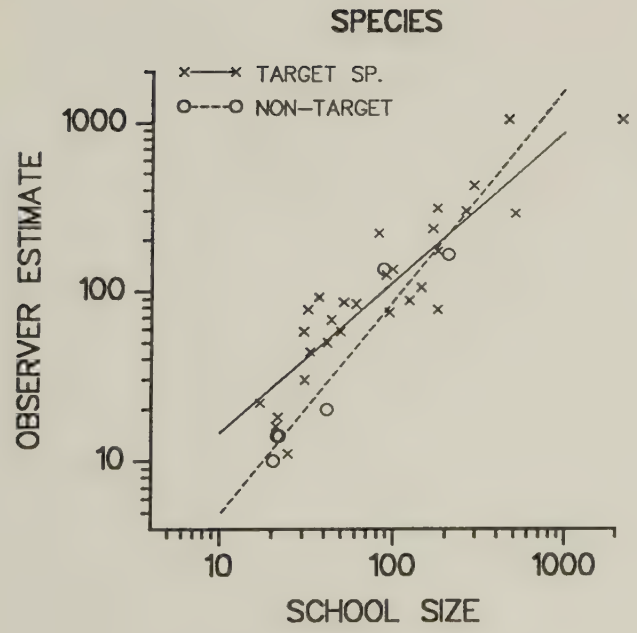


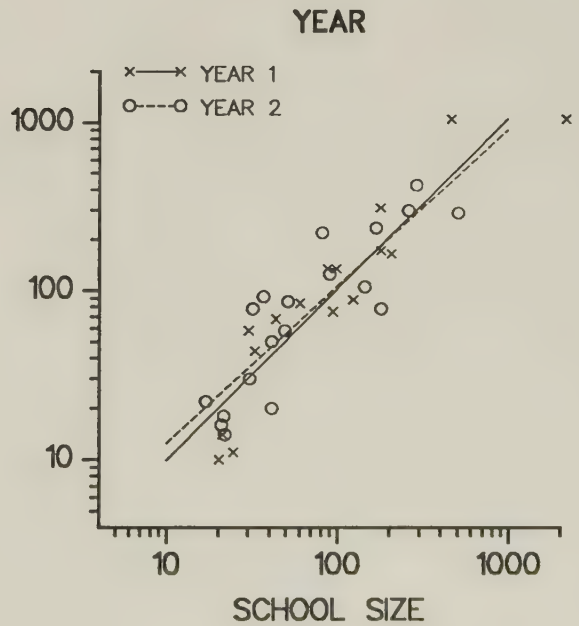
OBSERVER 8

N = 35



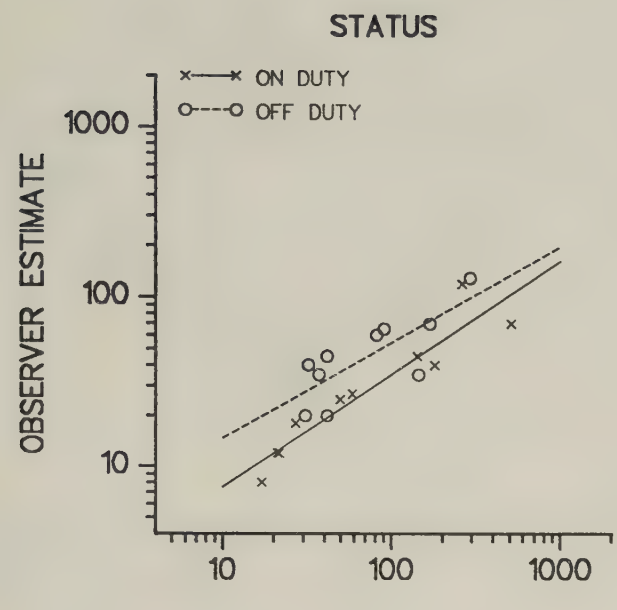


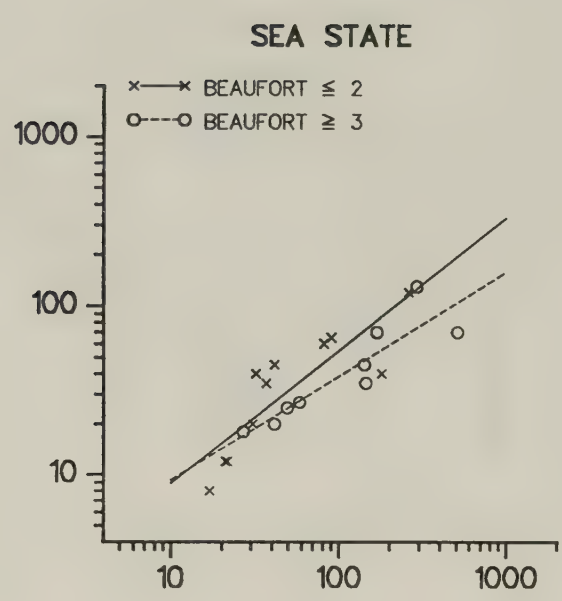


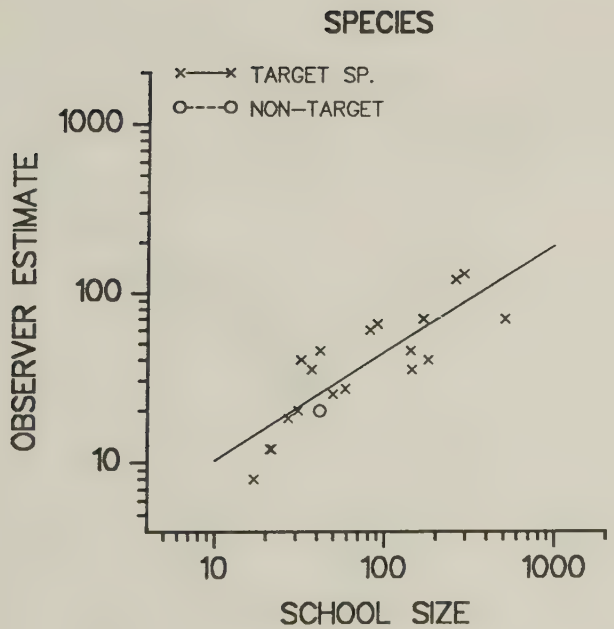


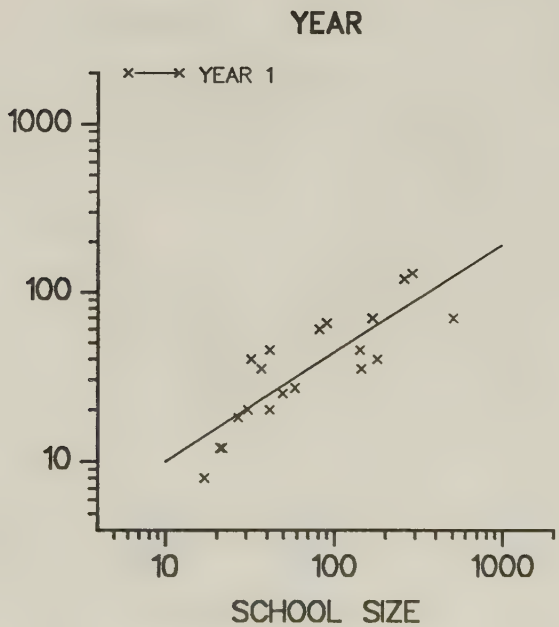
OBSERVER 9

N = 20



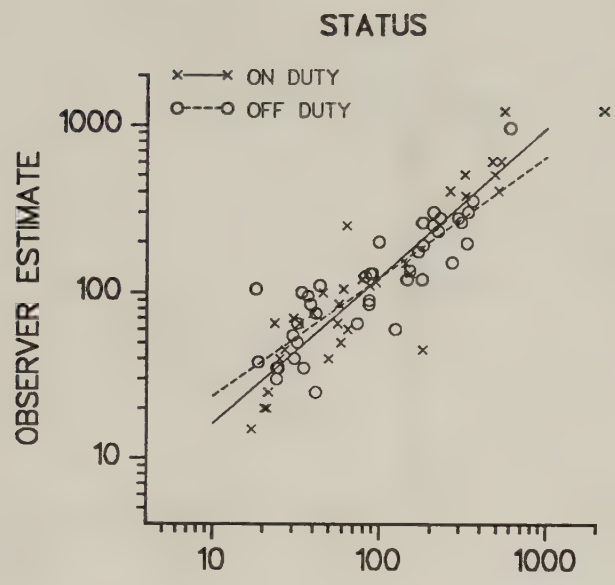


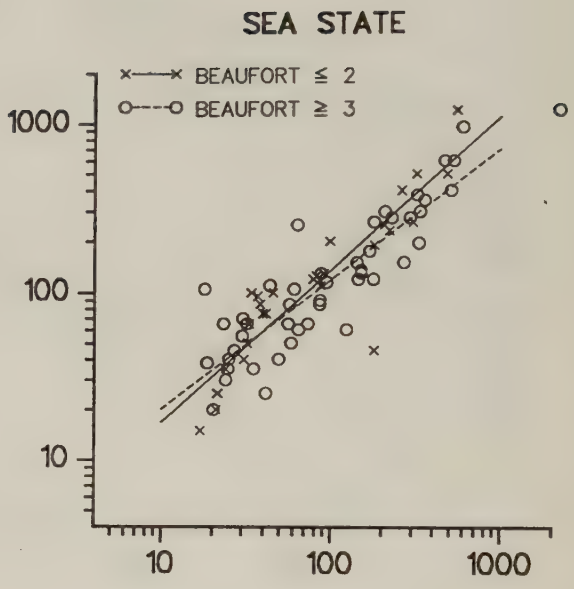


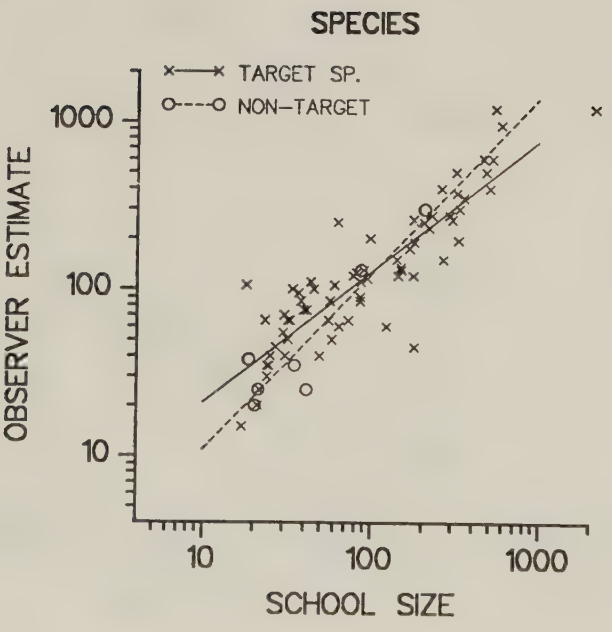


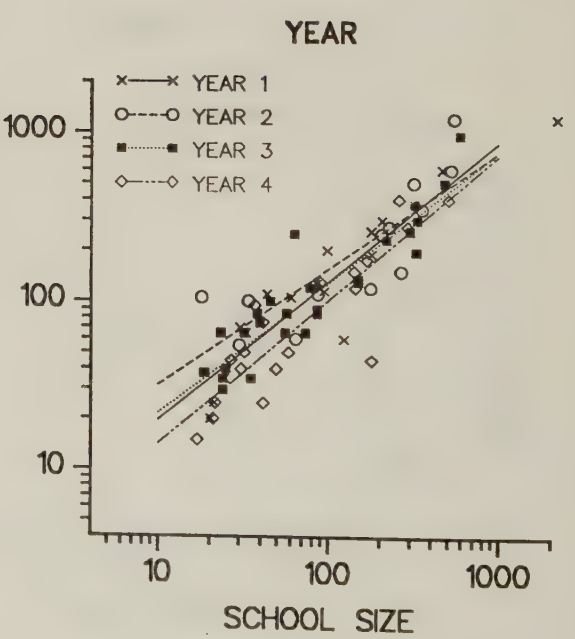
OBSERVER 10

N = 75



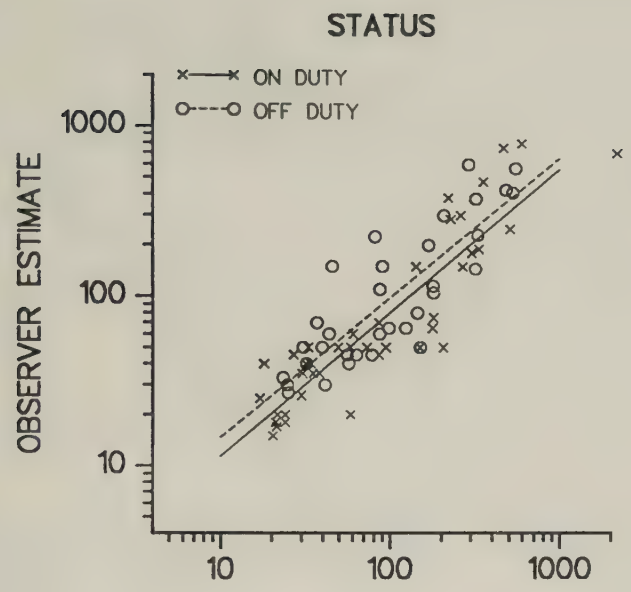


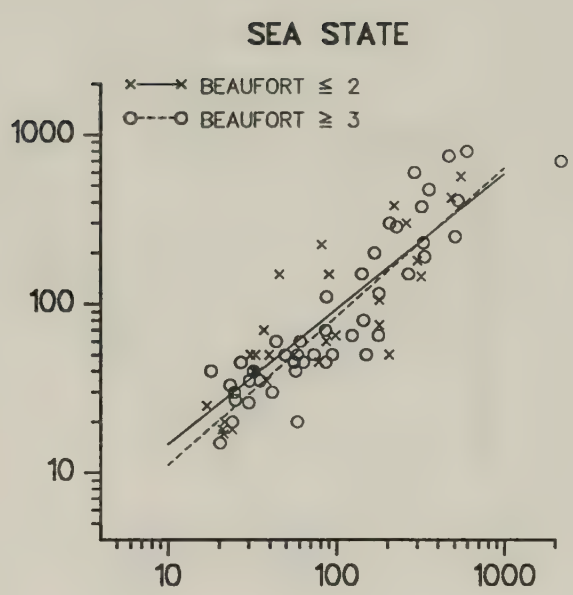


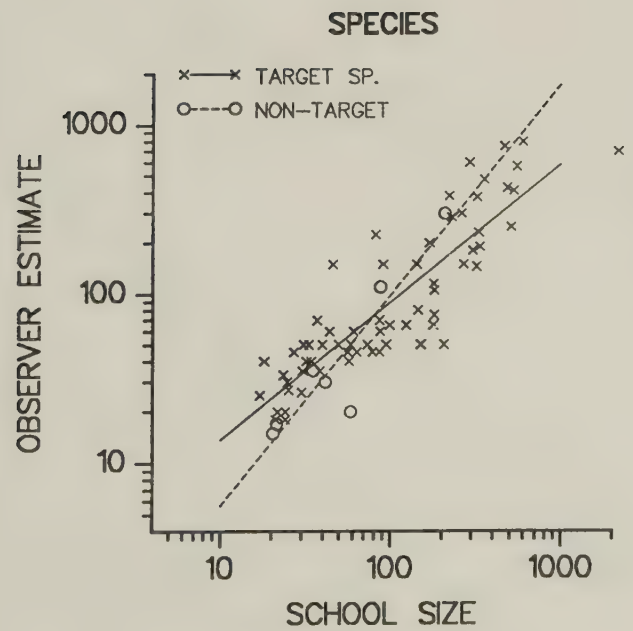


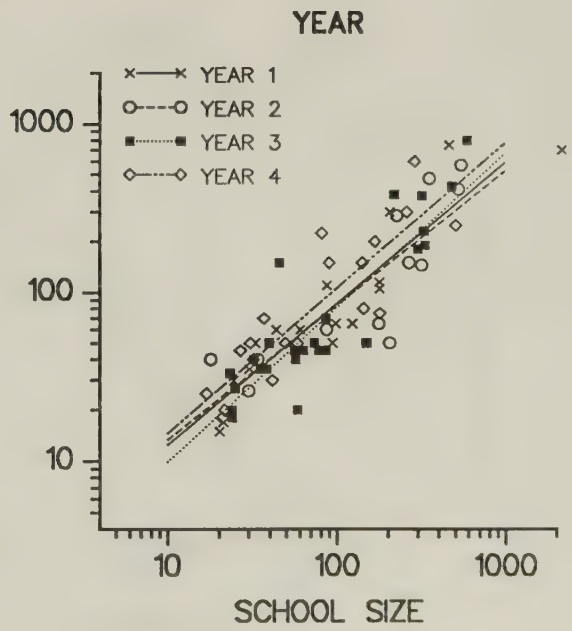
OBSERVER 11

N = 73



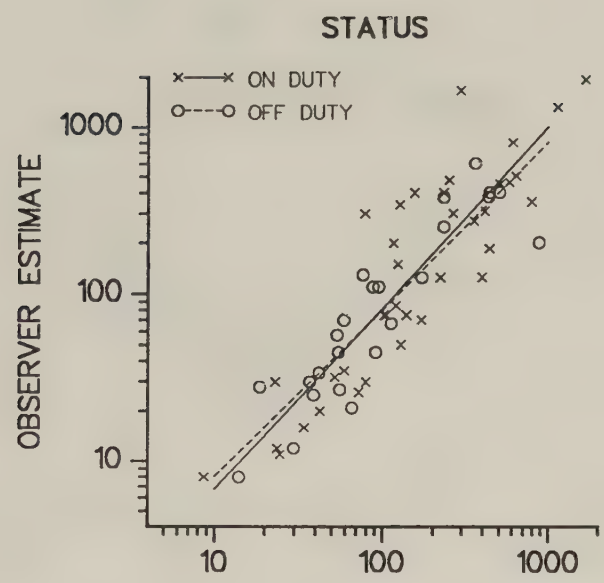


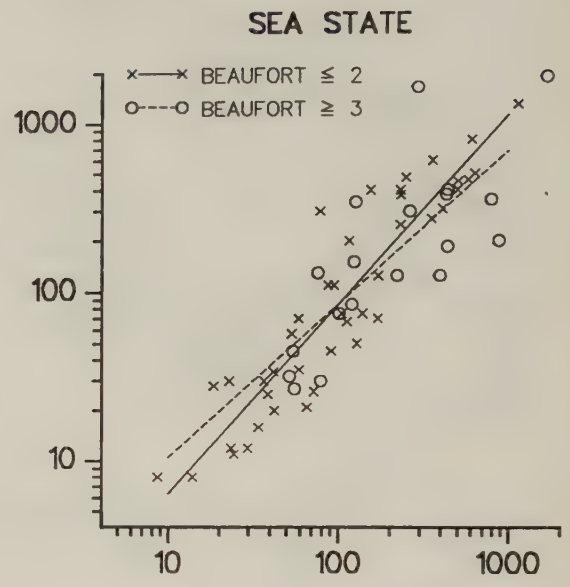


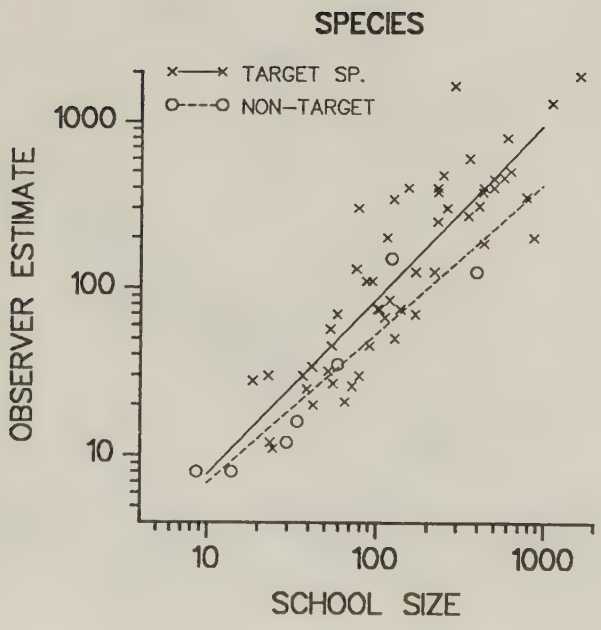


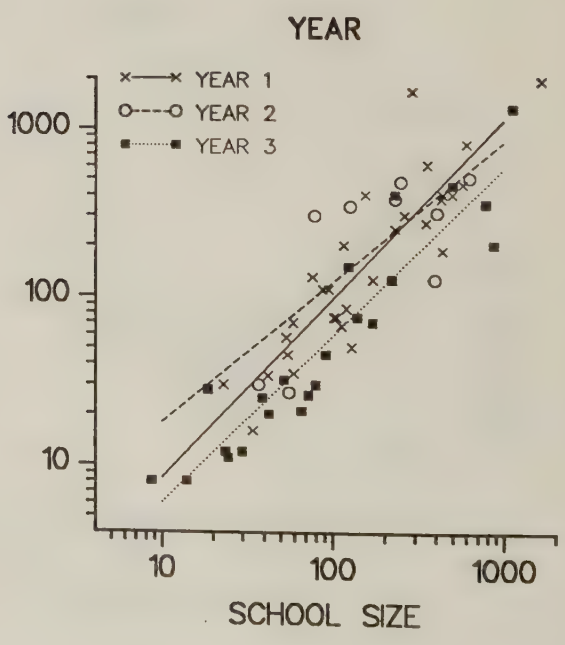
OBSERVER 12

N = 61



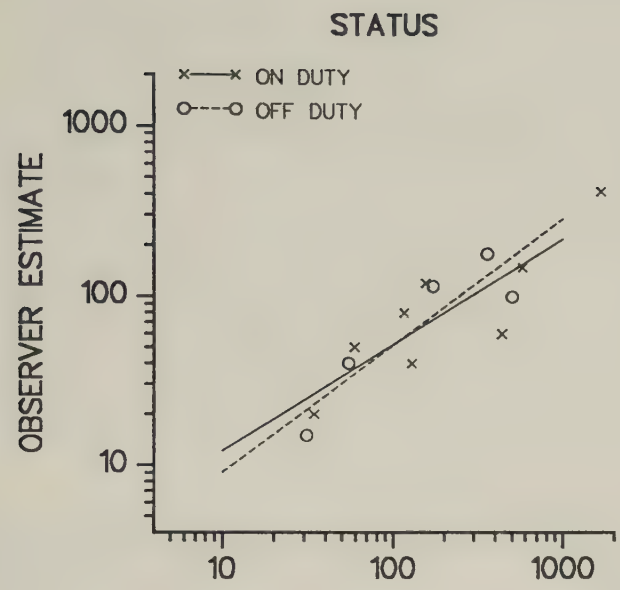


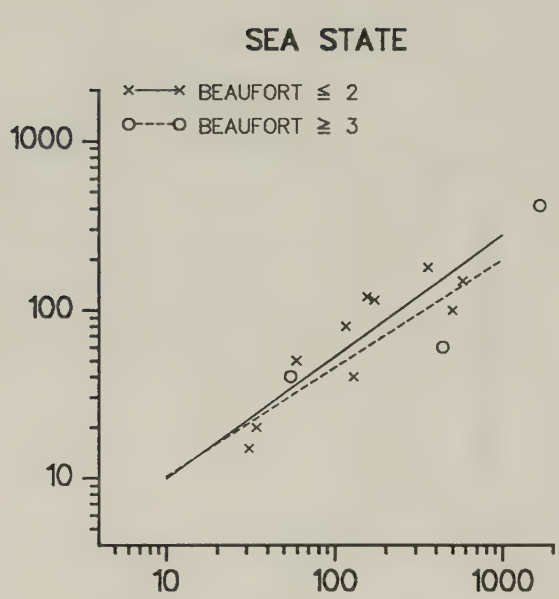


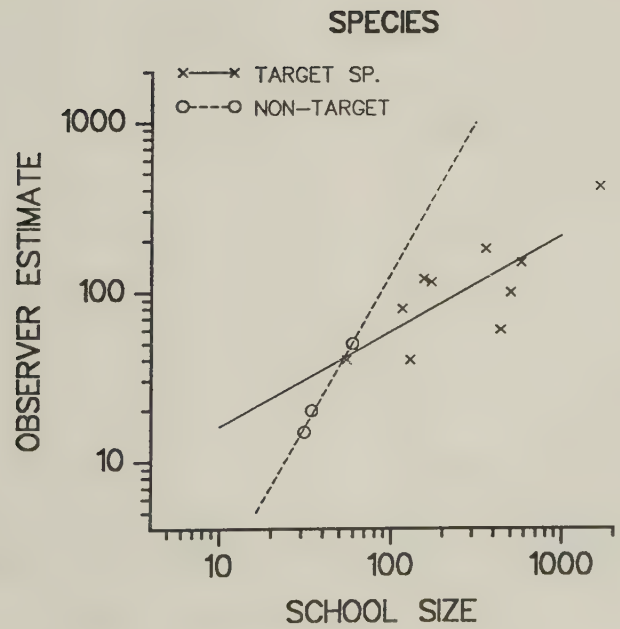


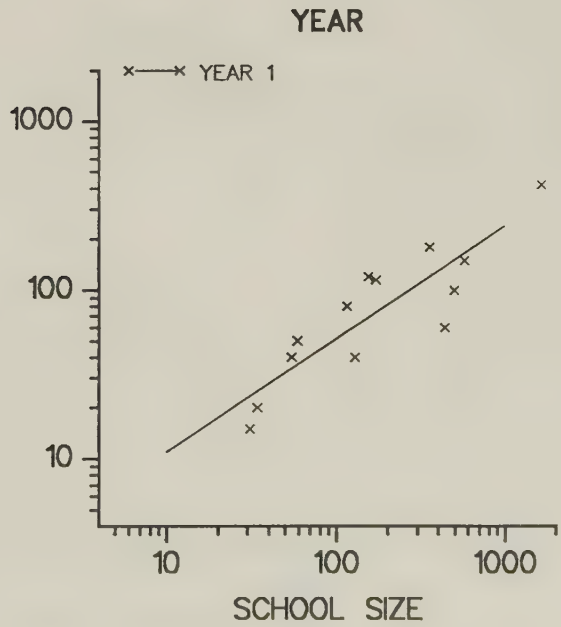
OBSERVER 13

N = 13



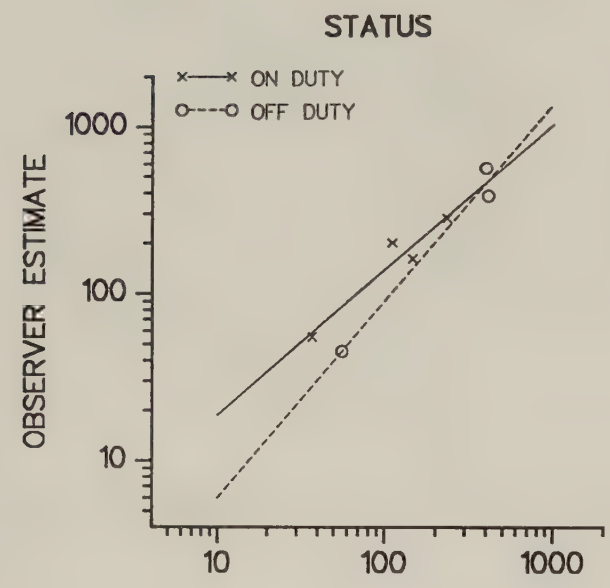


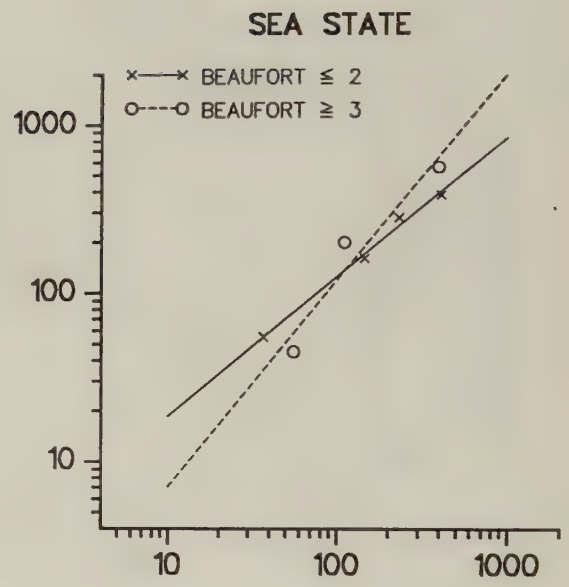


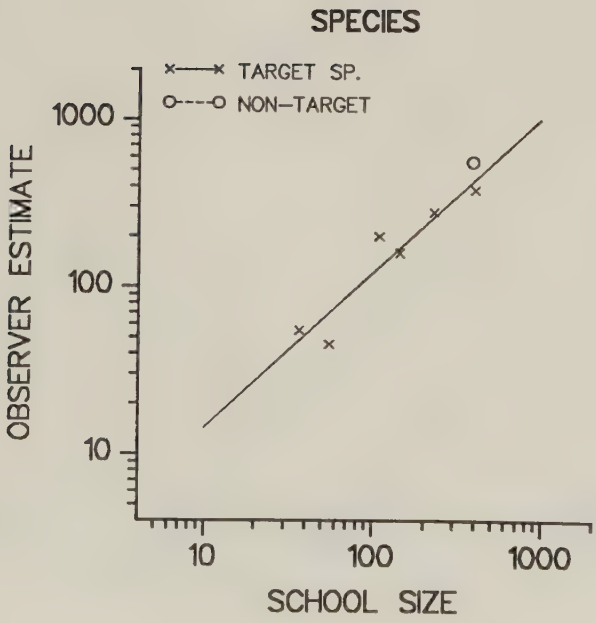


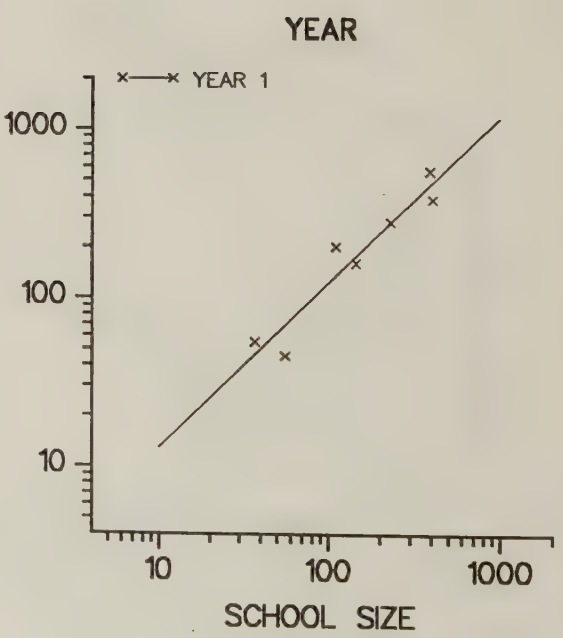
OBSERVER 14

N = 7



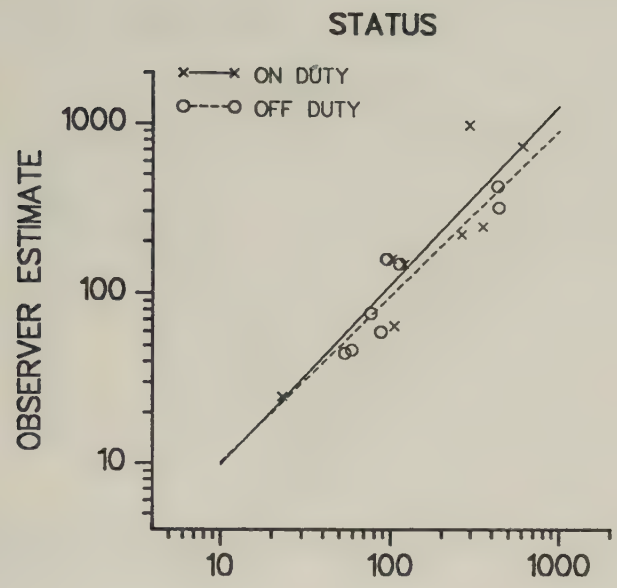


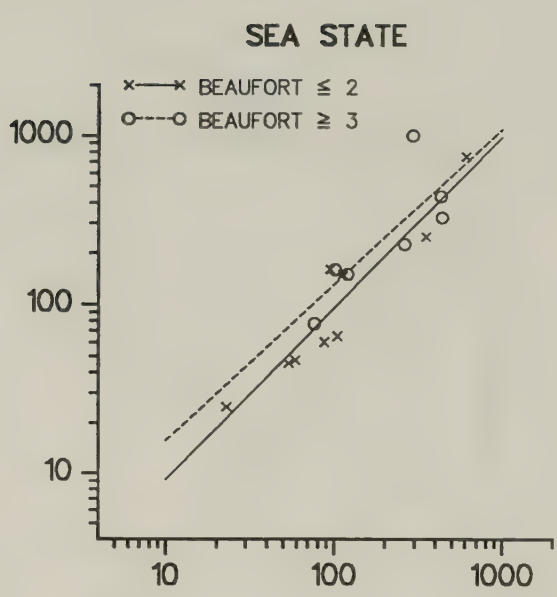


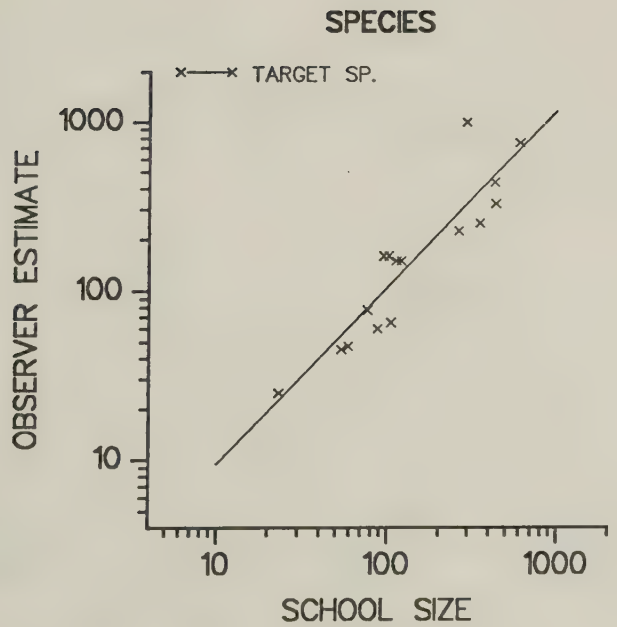


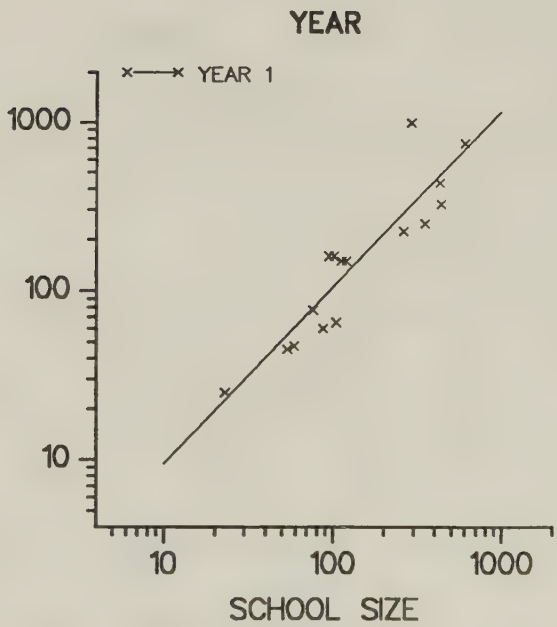
OBSERVER 15

N = 16



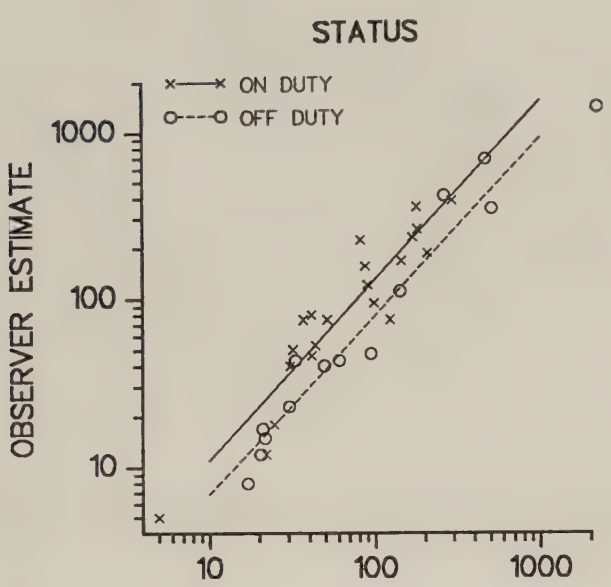


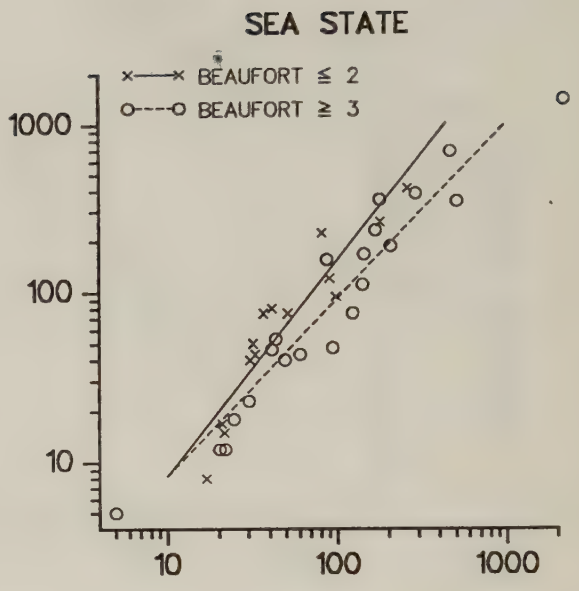


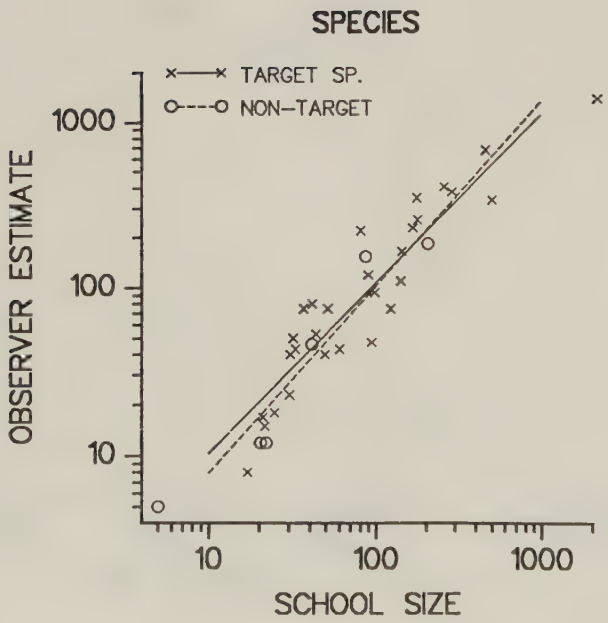


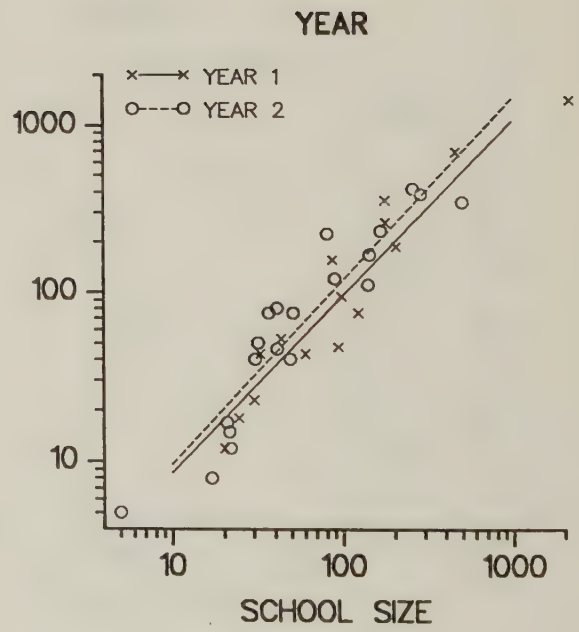
OBSERVER 16

N = 35



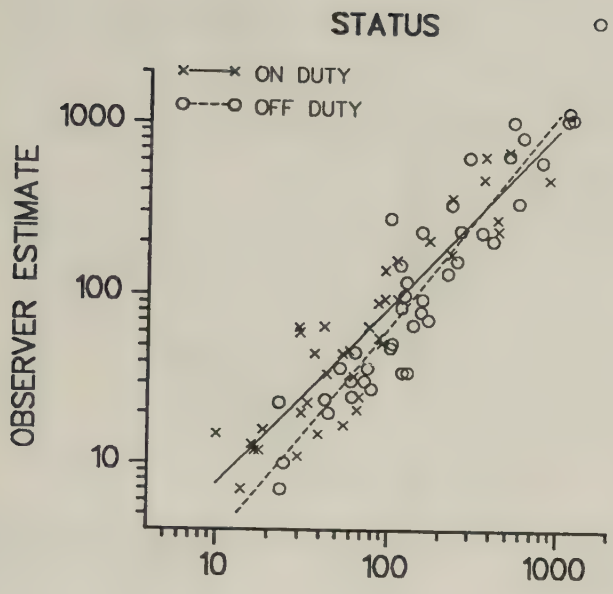


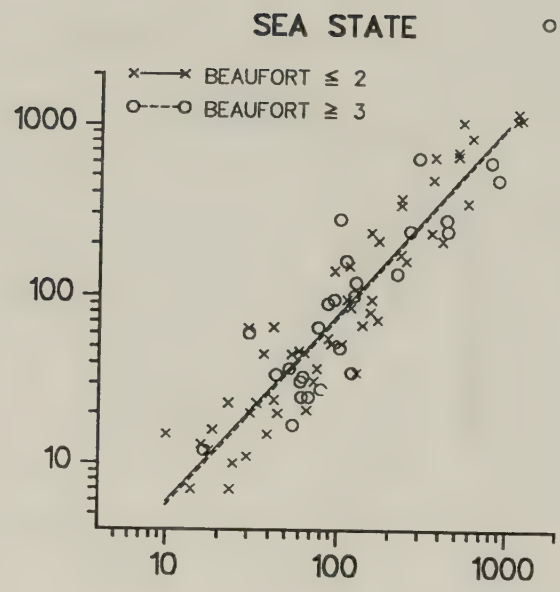


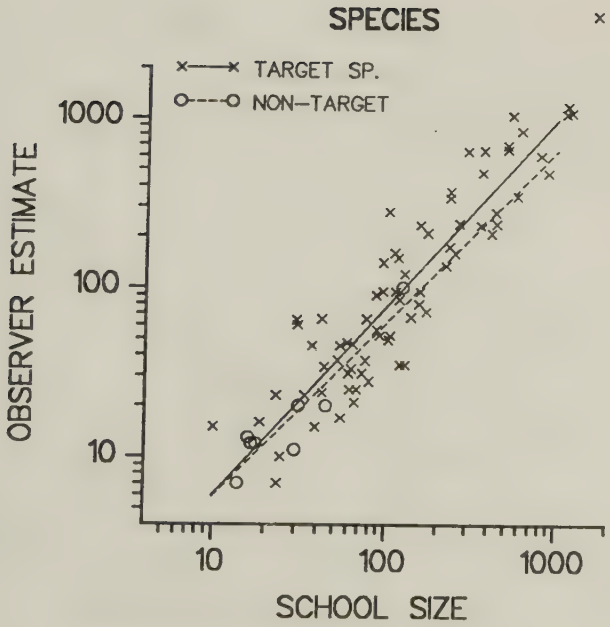


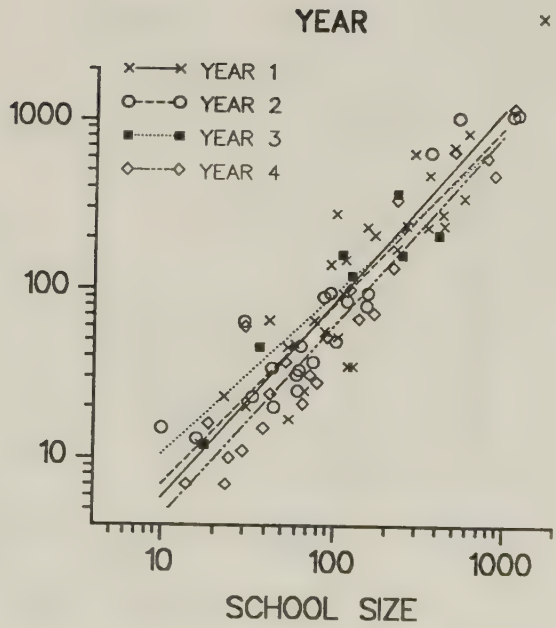
OBSERVER 17

N = 80



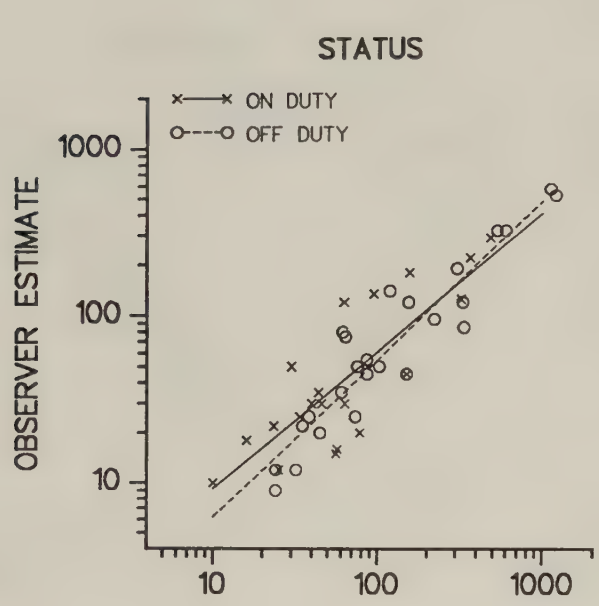


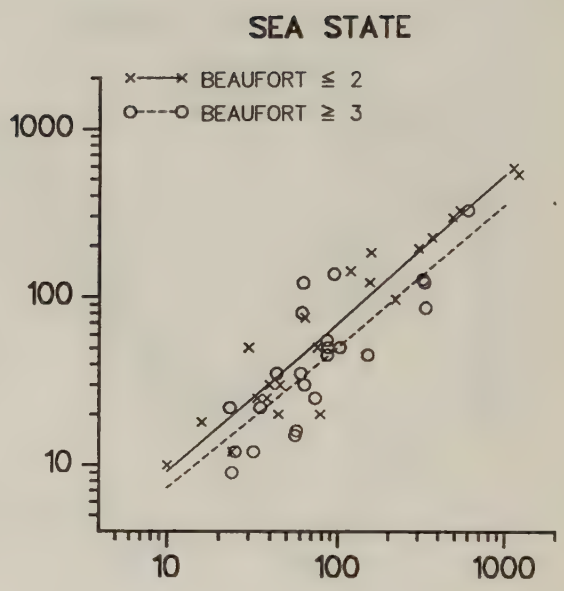


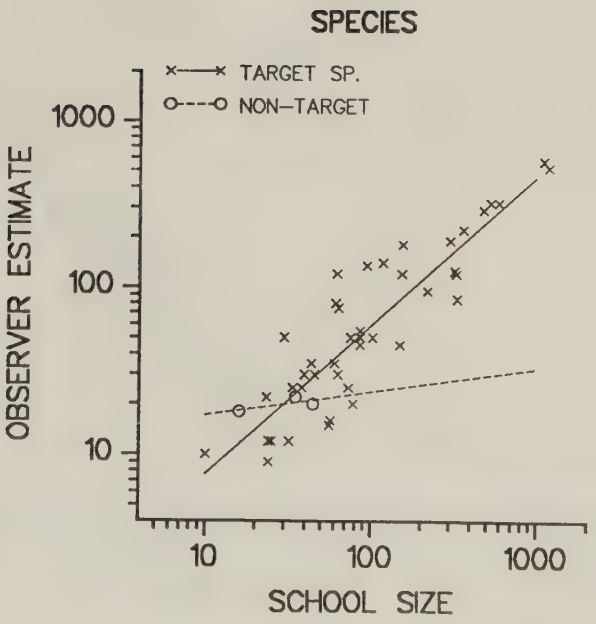


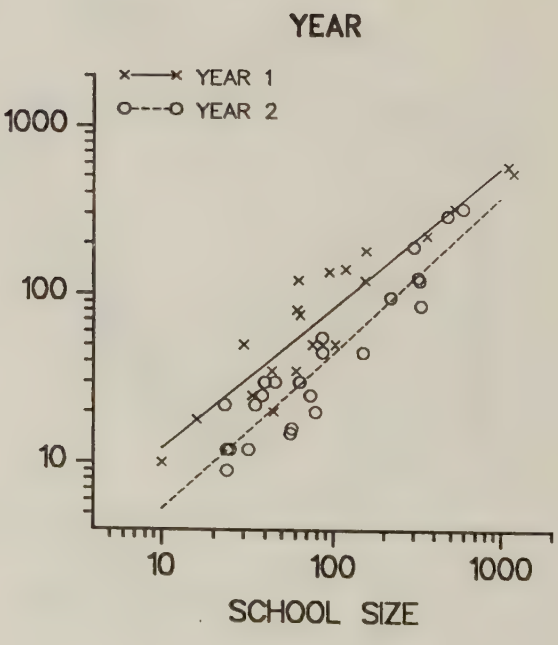
OBSERVER 18

N = 46



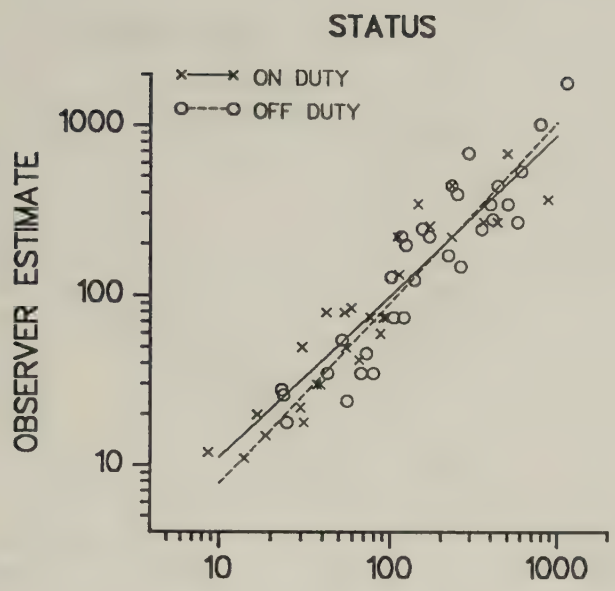


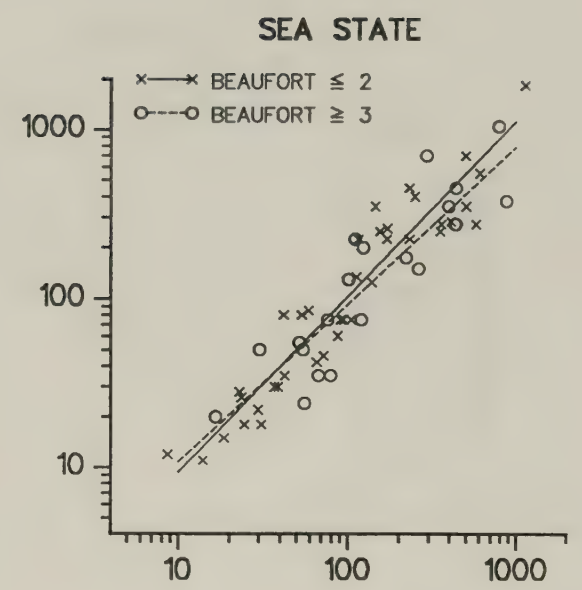


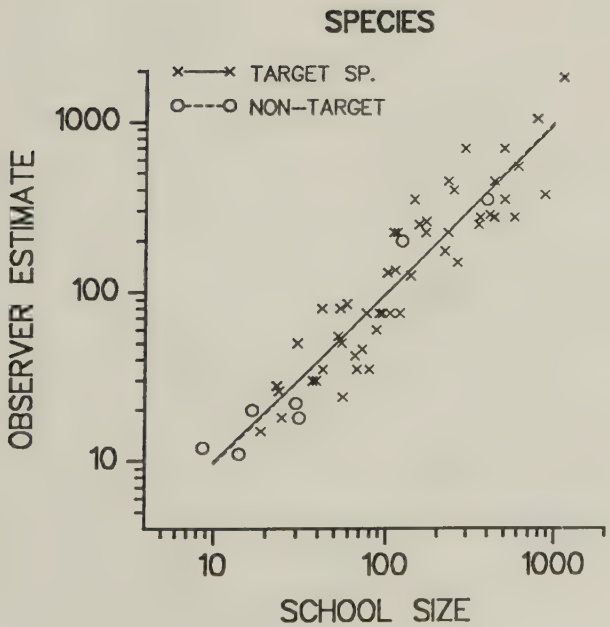


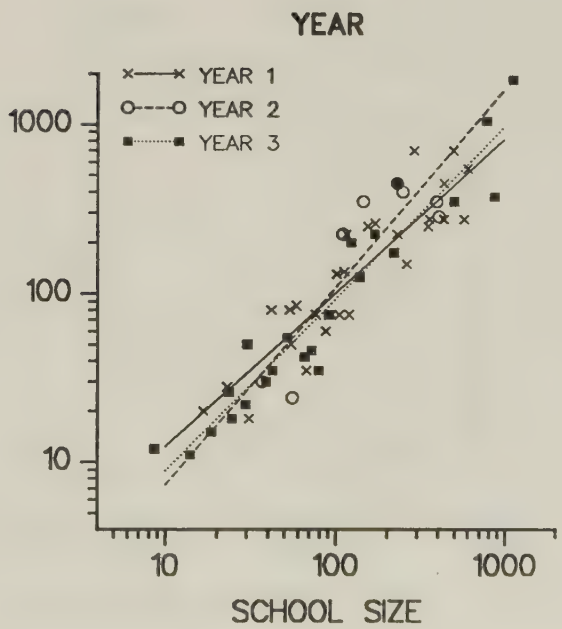
OBSERVER 19

N = 60



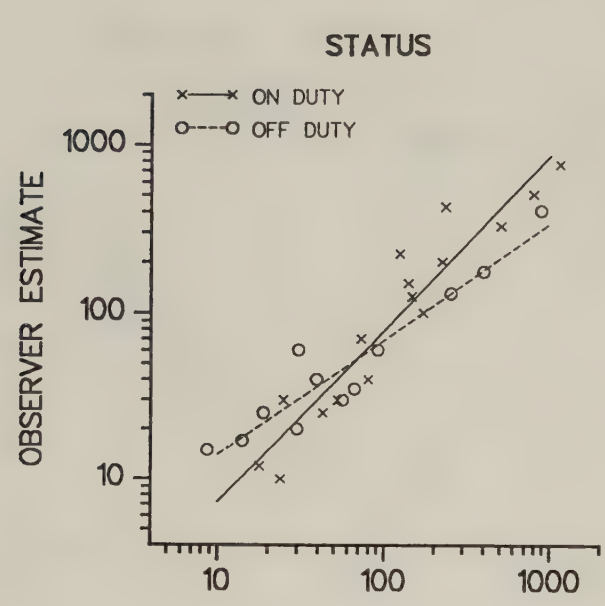


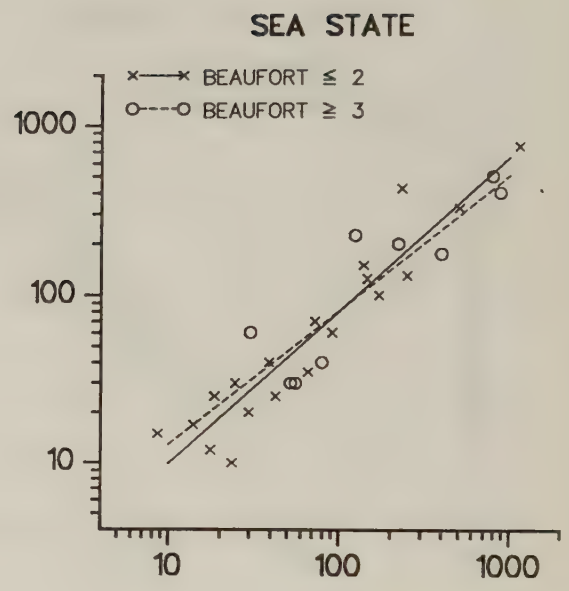


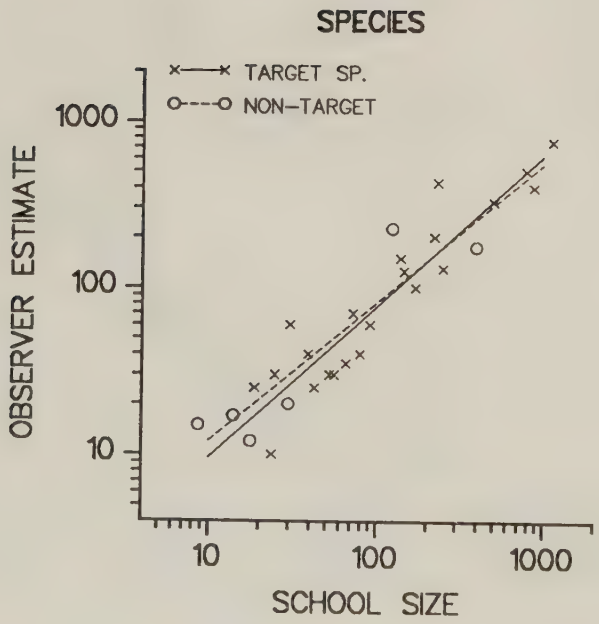


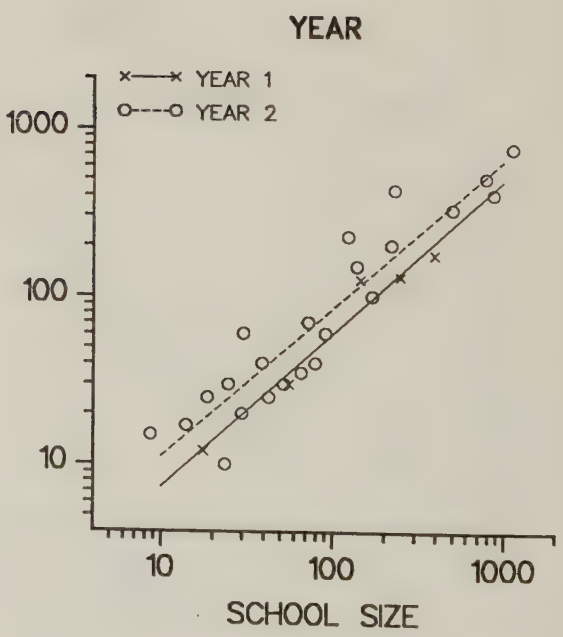
OBSERVER 20

N = 28



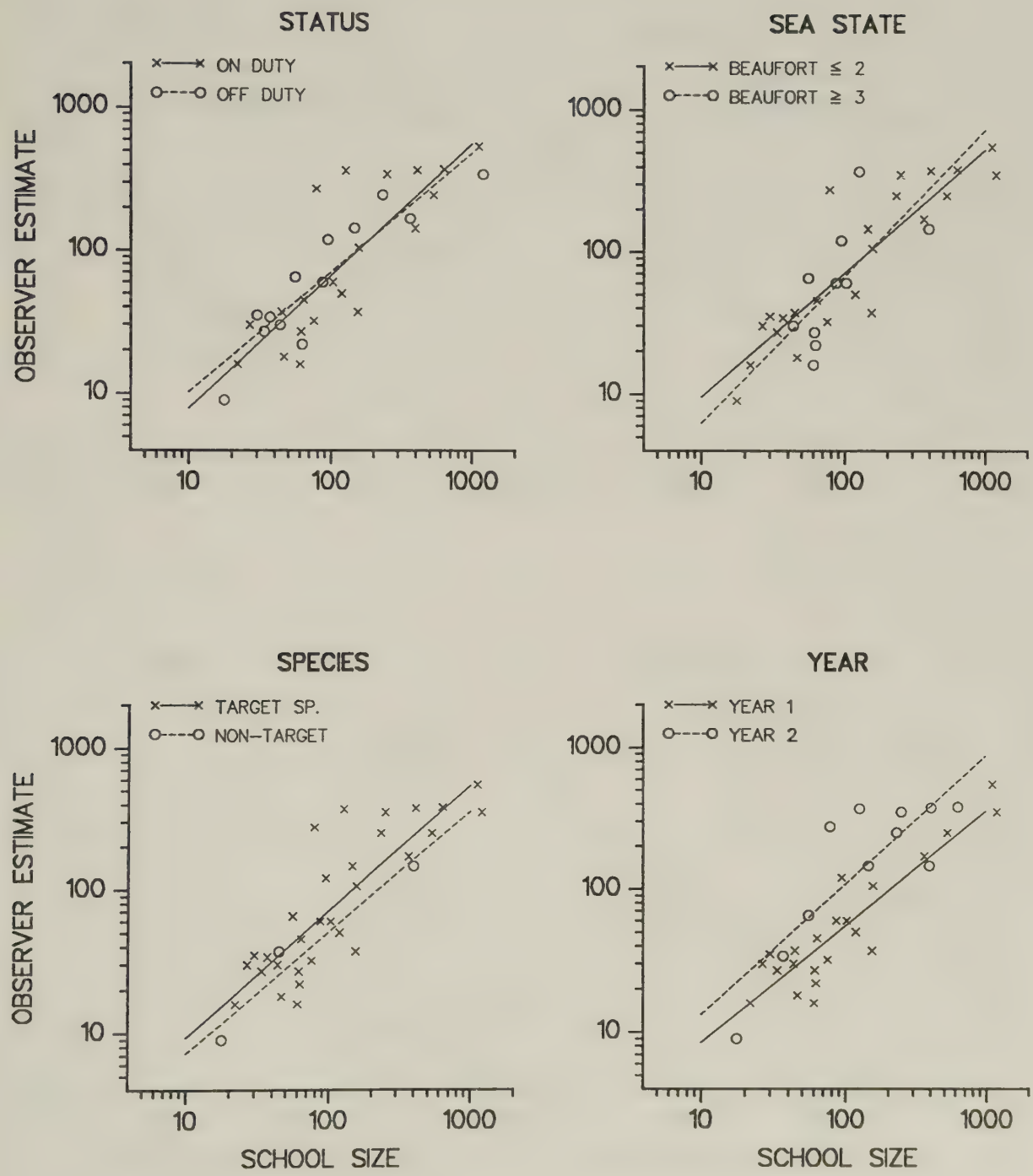






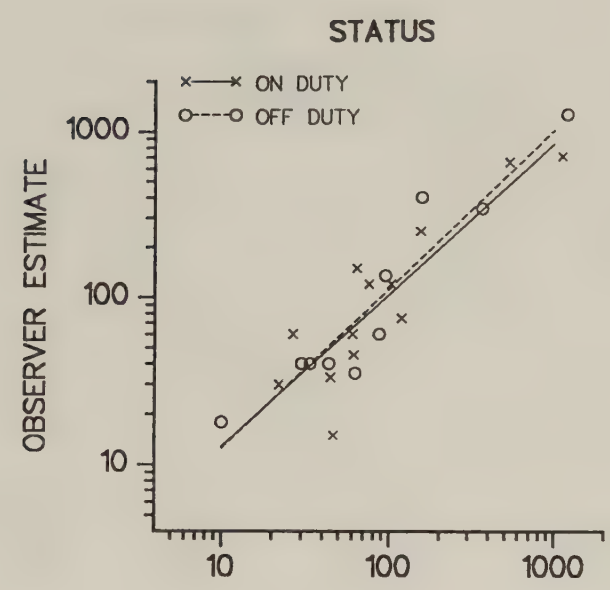
OBSERVER 21

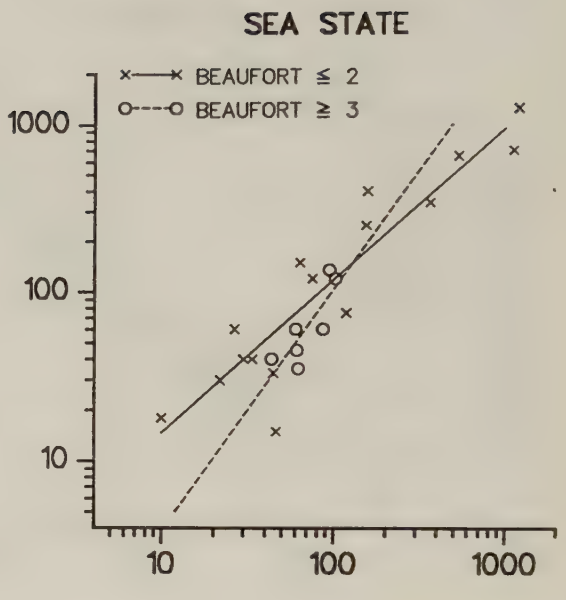
N = 33

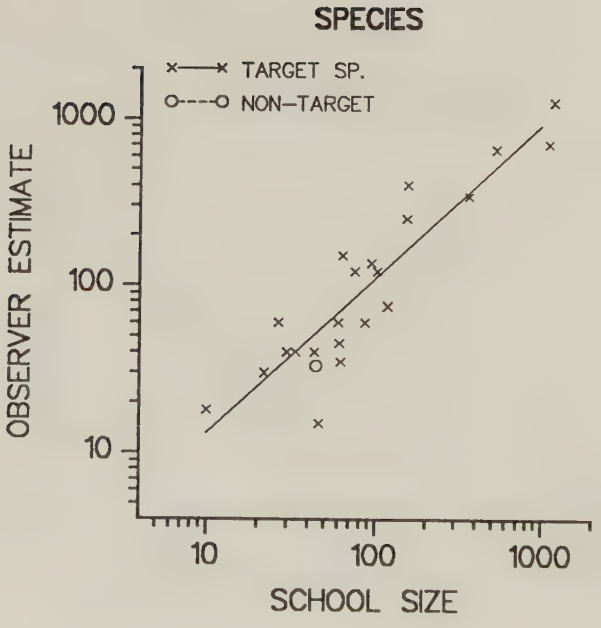


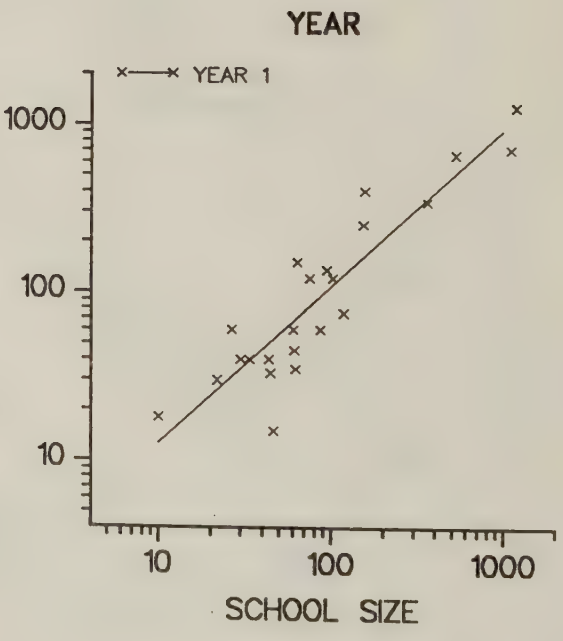
OBSERVER 22

N = 23



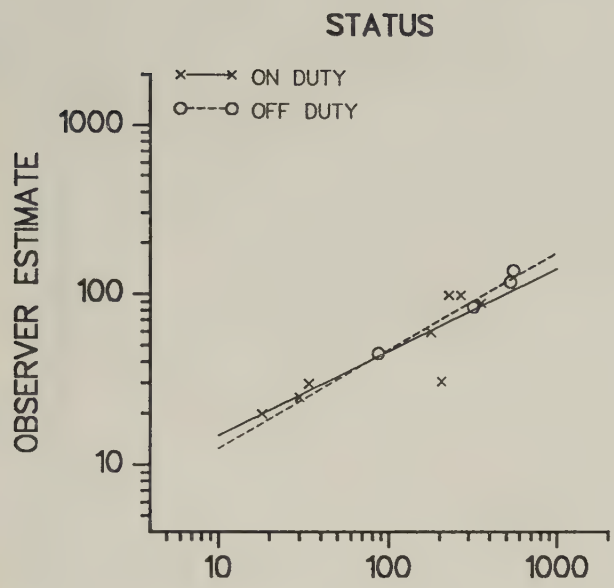


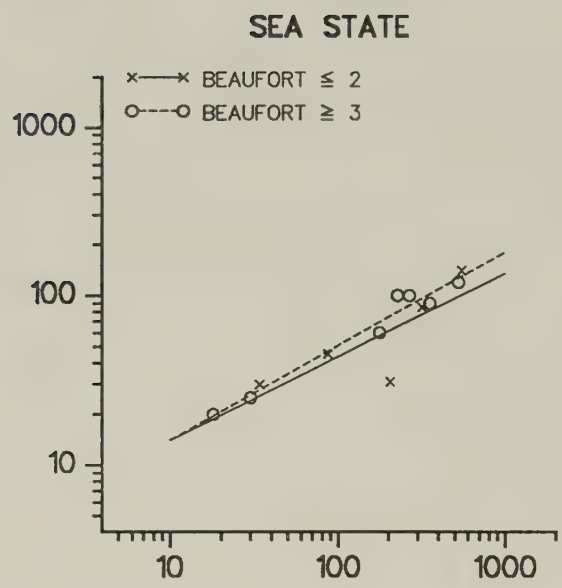


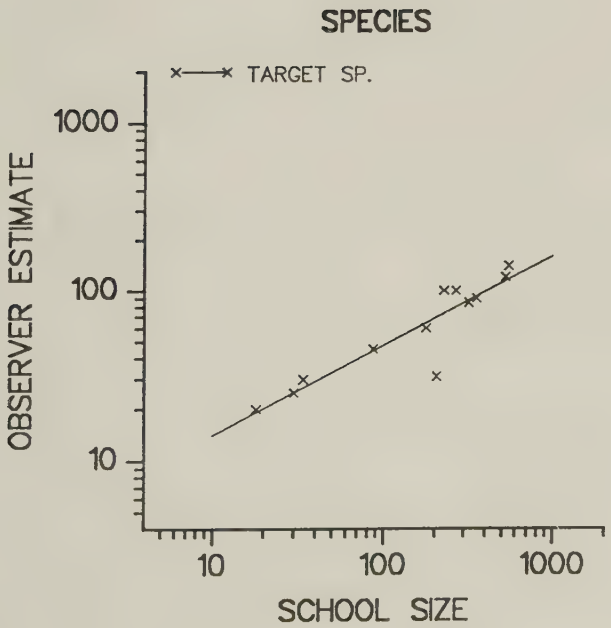


OBSERVER 23

N = 12







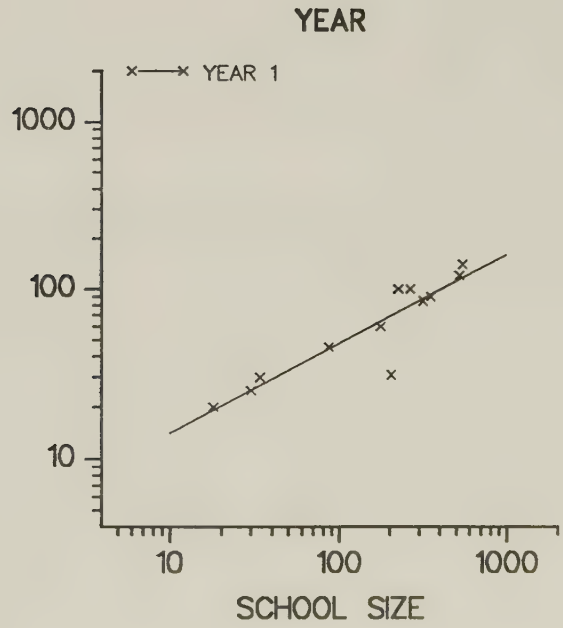
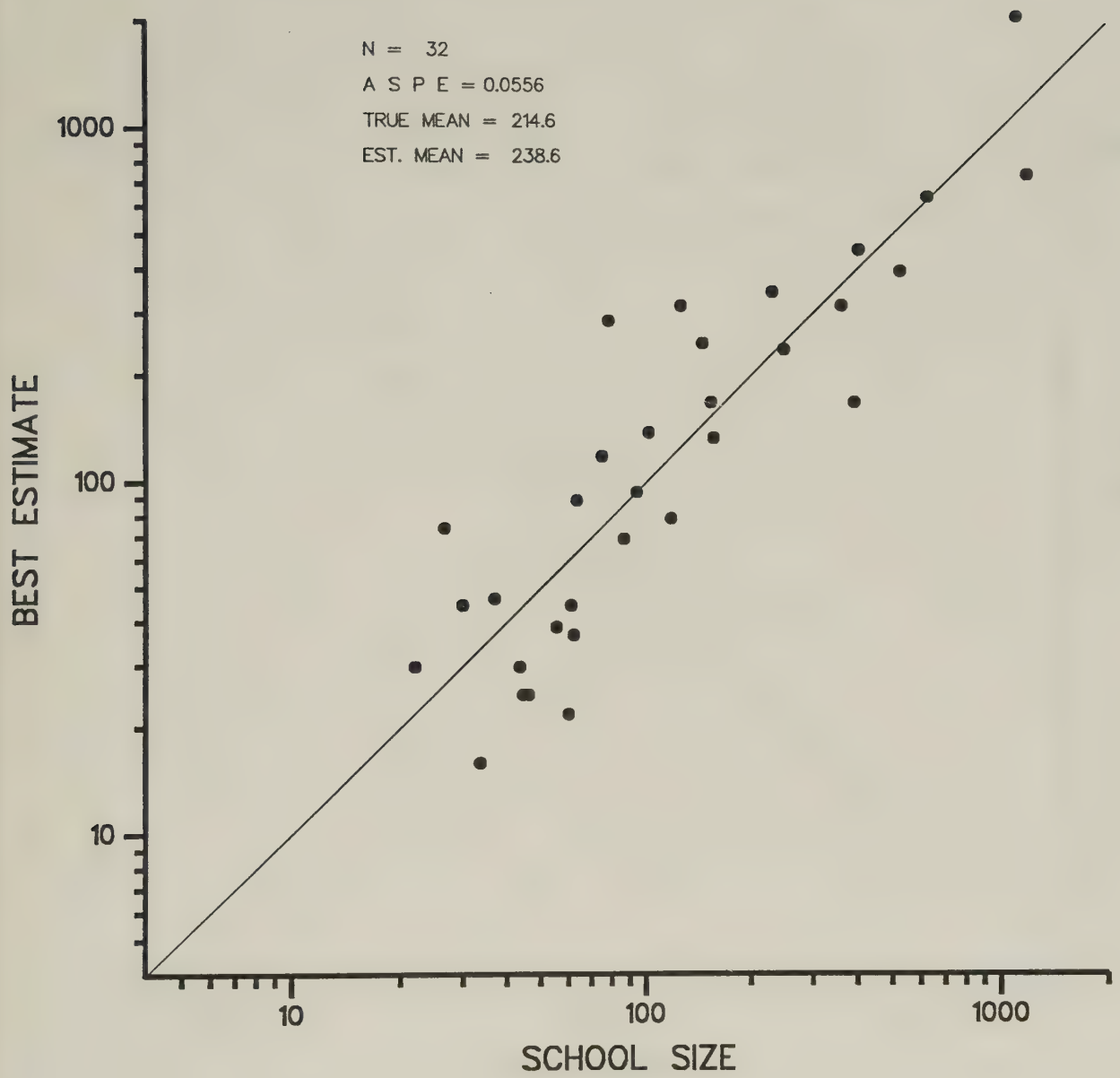


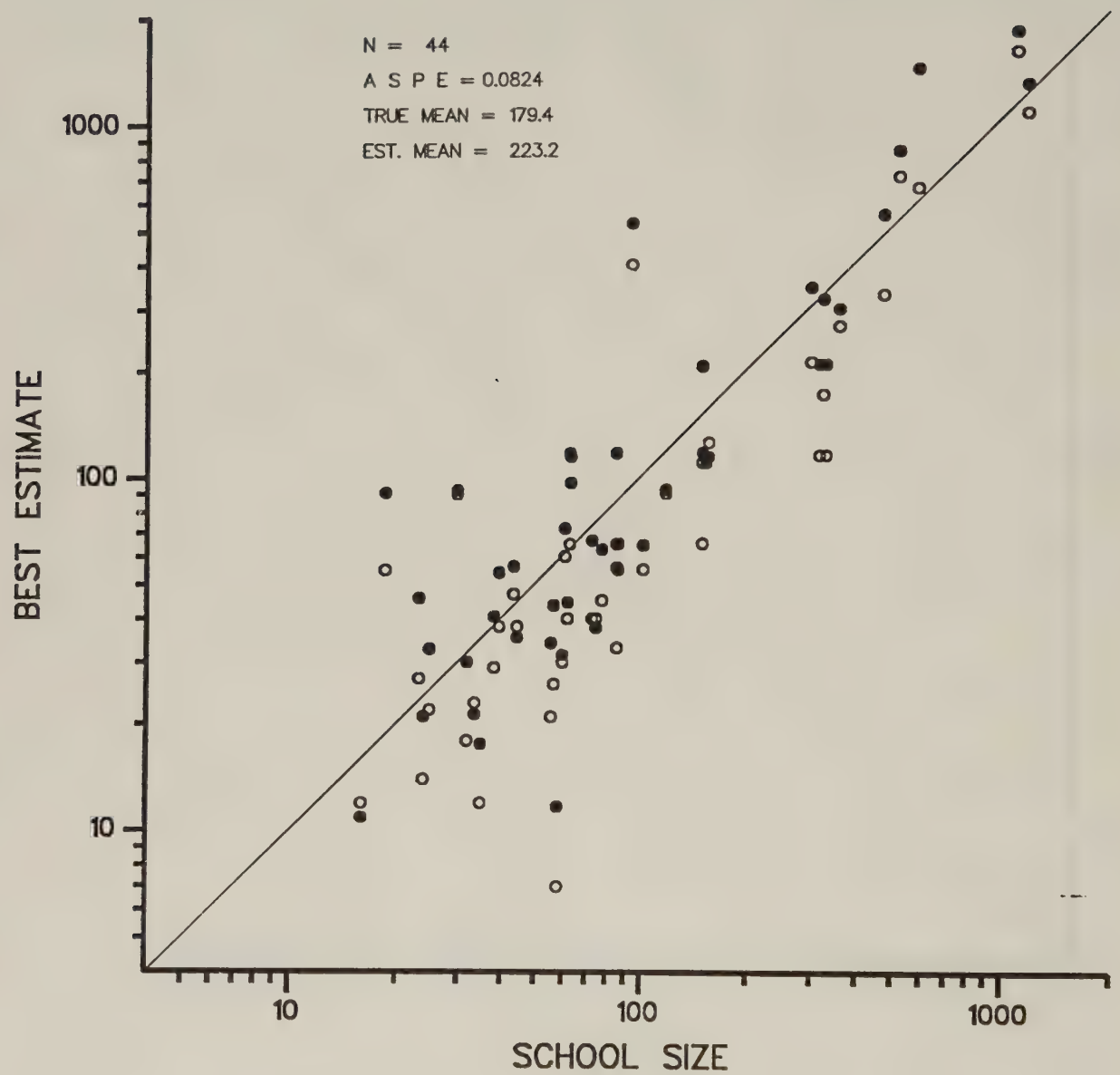
Fig. 2: Observer estimates of dolphin school size plotted against known school size using the best procedure for each fo the 23 observers in this study (filled symbols). For the 10 observers whose estimates were improved by calibration, their unadjusted best estimates are also plotted as open symbols. The line in each graph is the 1:1 line, not a regression line.

.FOR

OBSERVER 1

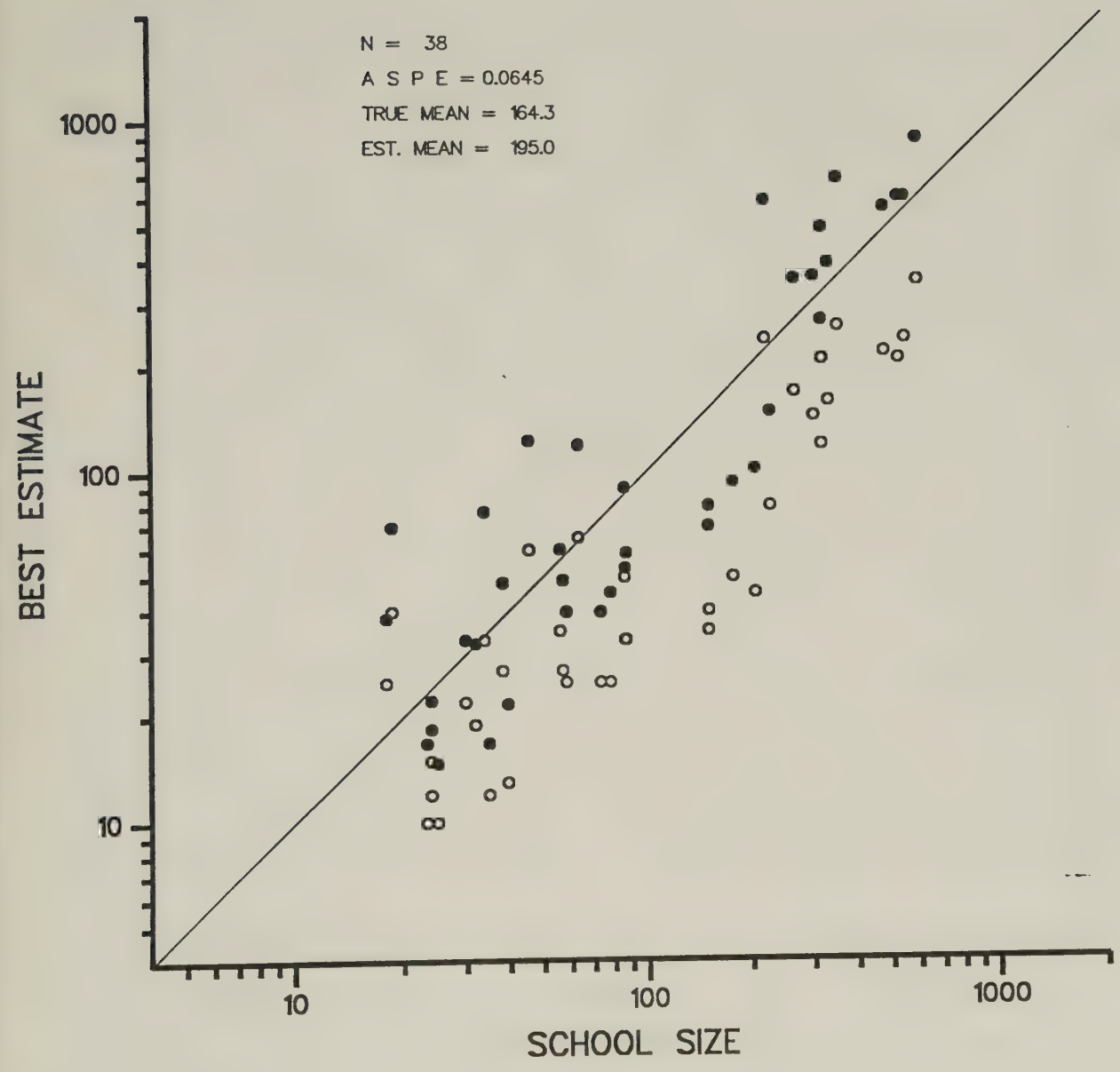


OBSERVER 2

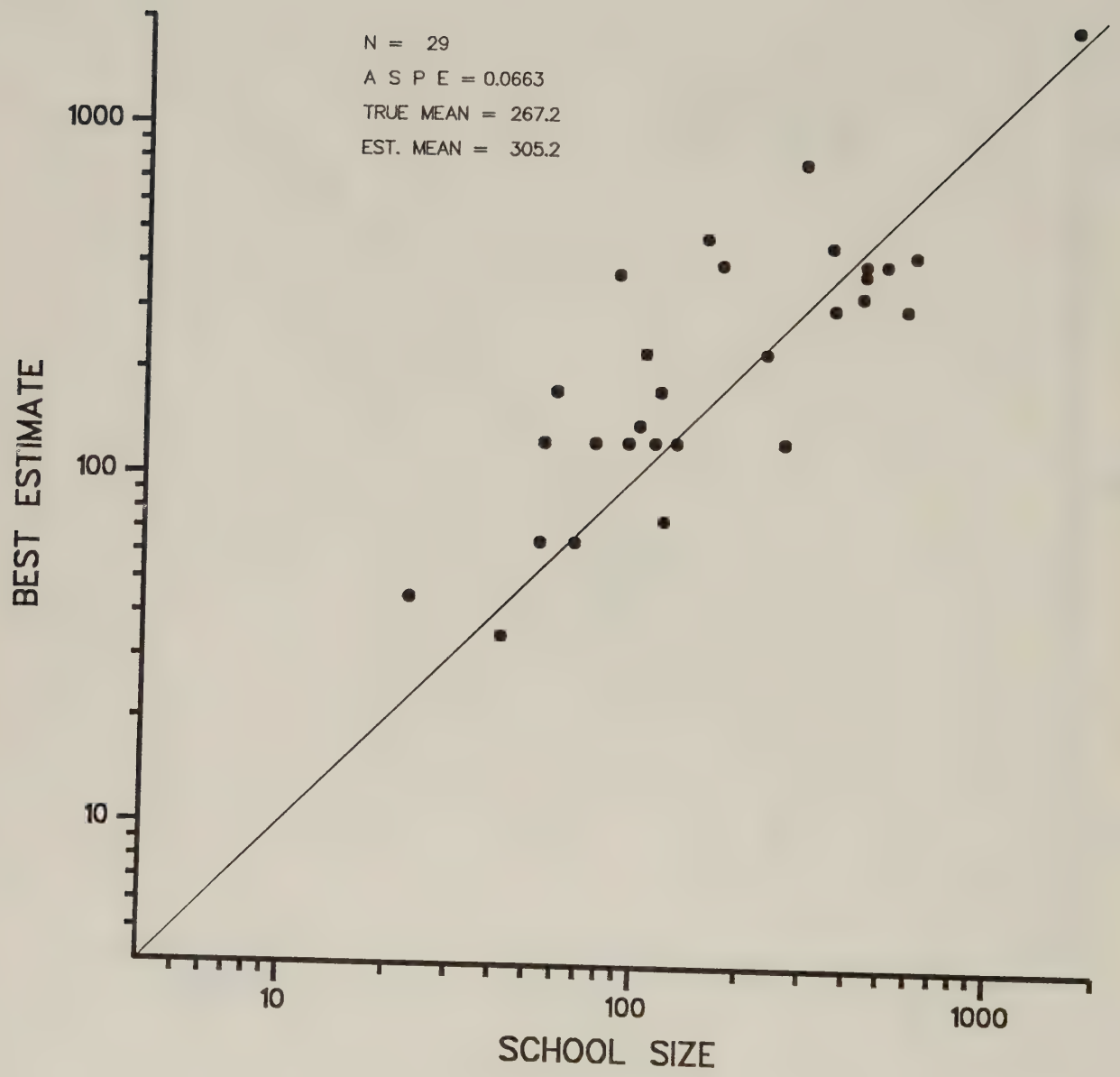

N = 44
A S P E = 0.0824
TRUE MEAN = 179.4
EST. MEAN = 223.2
BEST ESTIMATE
1000
100
10
10
100
1000
SCHOOL SIZE

FOR

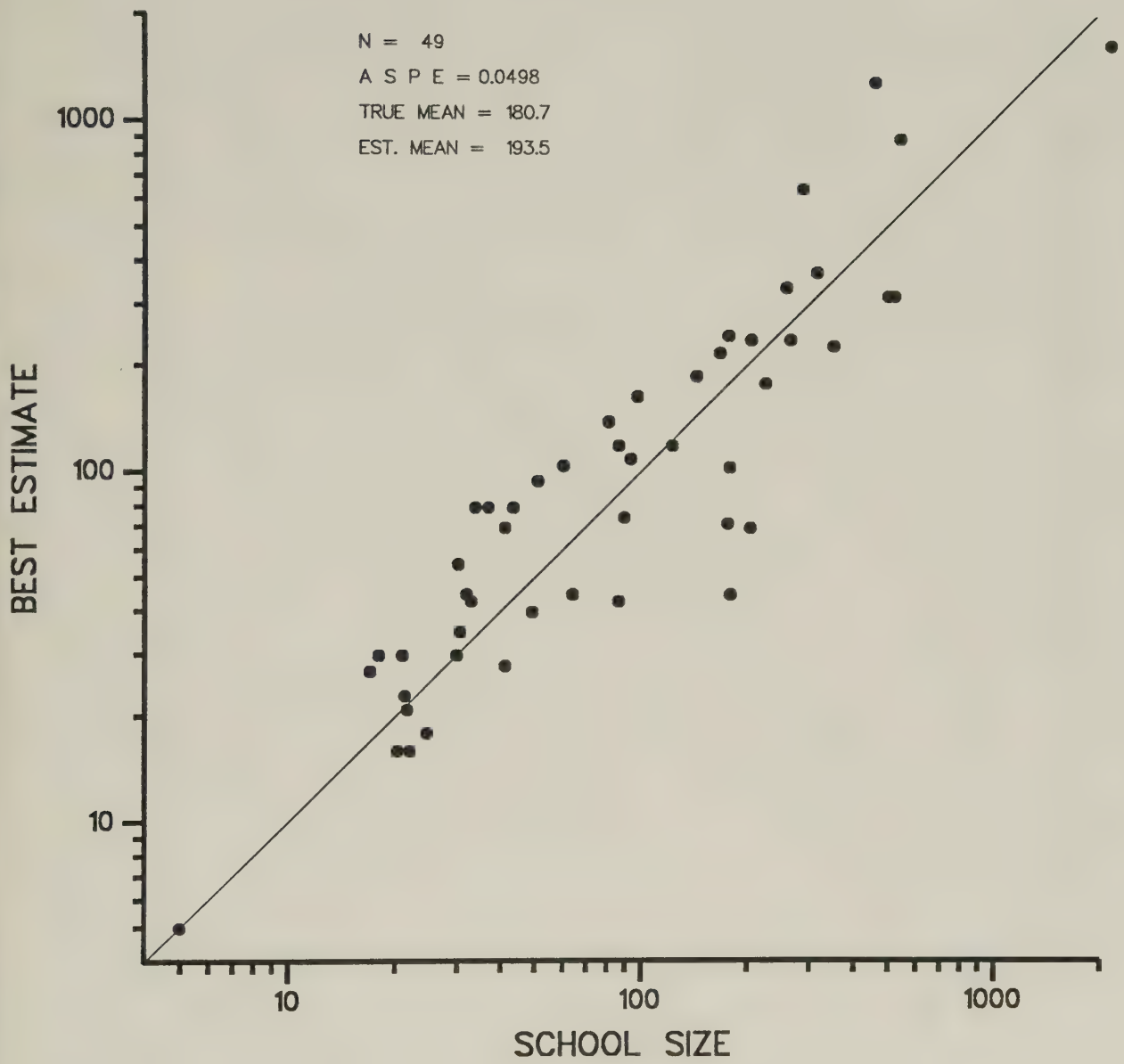
OBSERVER 3



OBSERVER 4

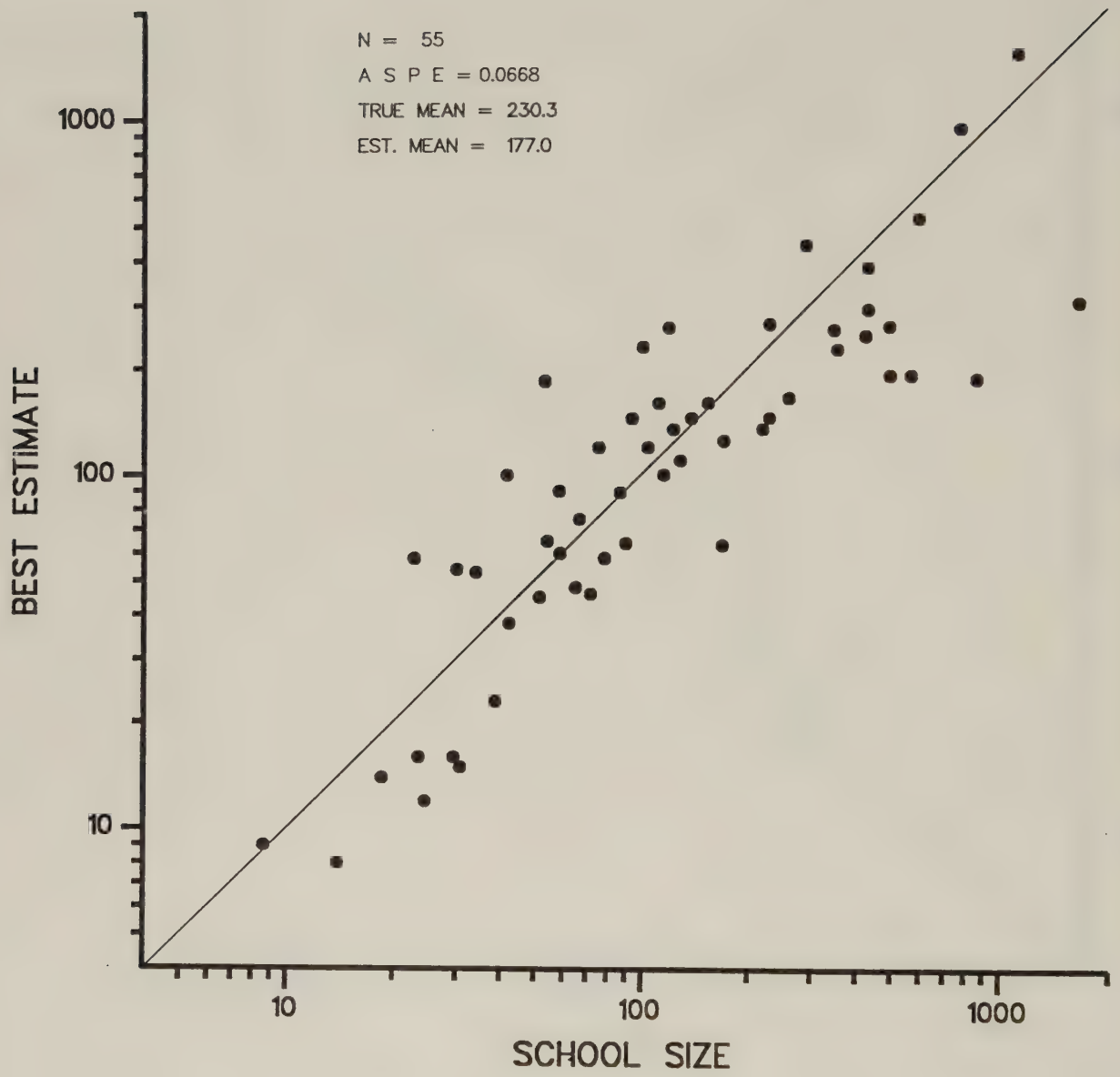

N = 29
A S P E = 0.0663
TRUE MEAN = 267.2
EST. MEAN = 305.2
1000
100
10
BEST ESTIMATE
10
100
1000
SCHOOL SIZE

OBSERVER 5


N = 49
A S P E = 0.0498
TRUE MEAN = 180.7
EST. MEAN = 193.5
1000
100
10
BEST ESTIMATE
10
100
1000
SCHOOL SIZE

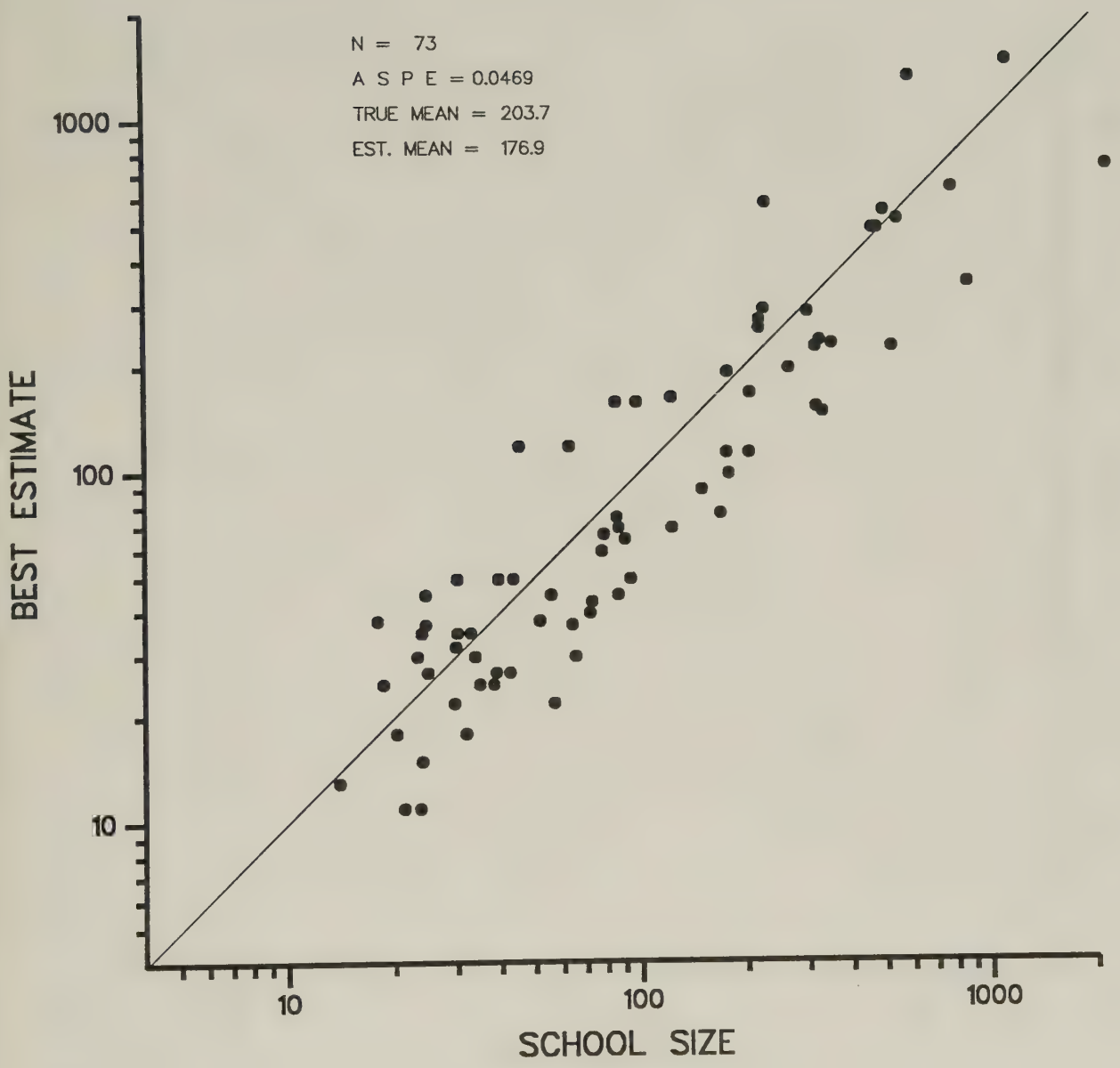
SKOOLWT.FOR
12/18/91

OBSERVER 6



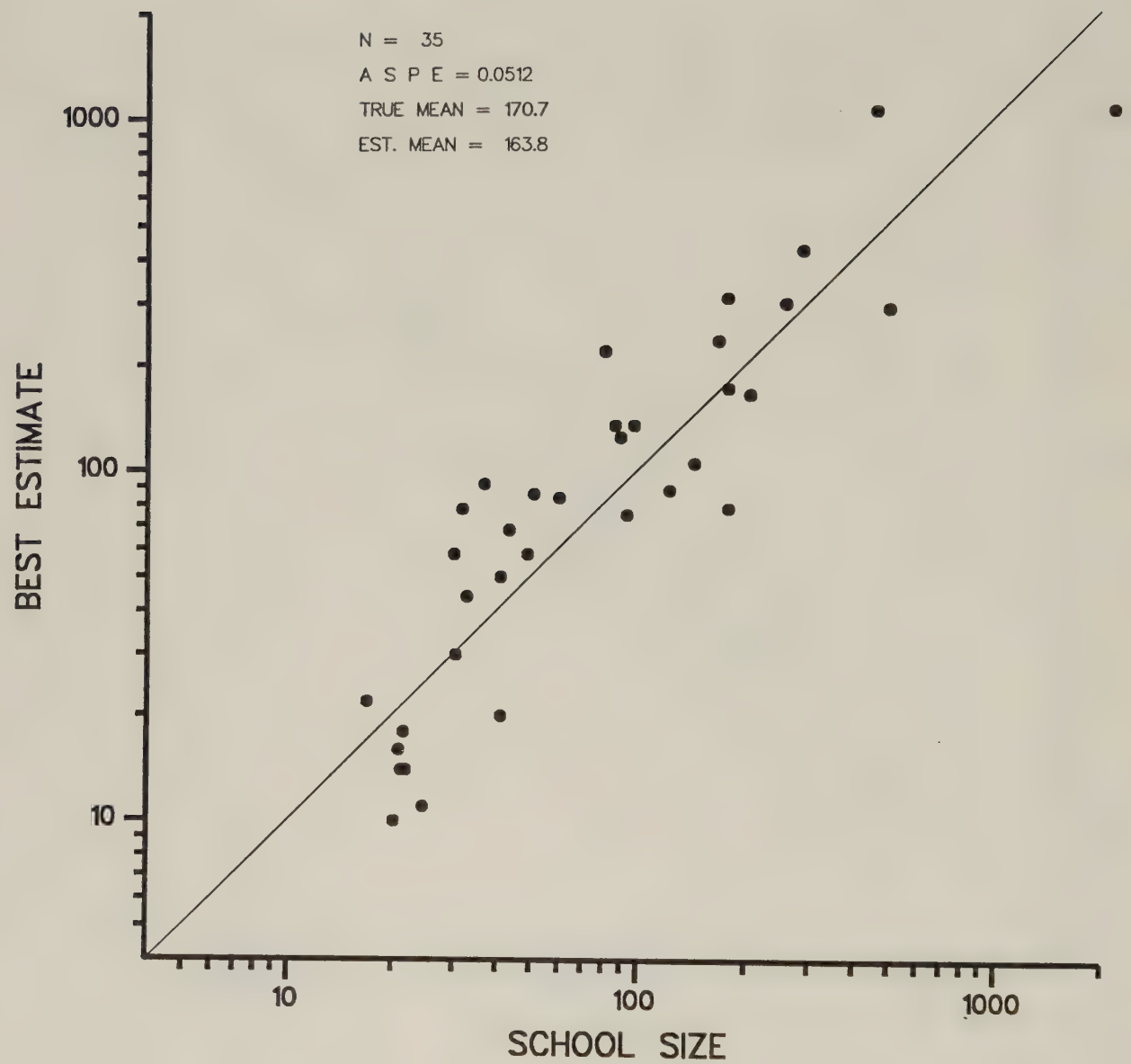
T.FOR
91

OBSERVER 7

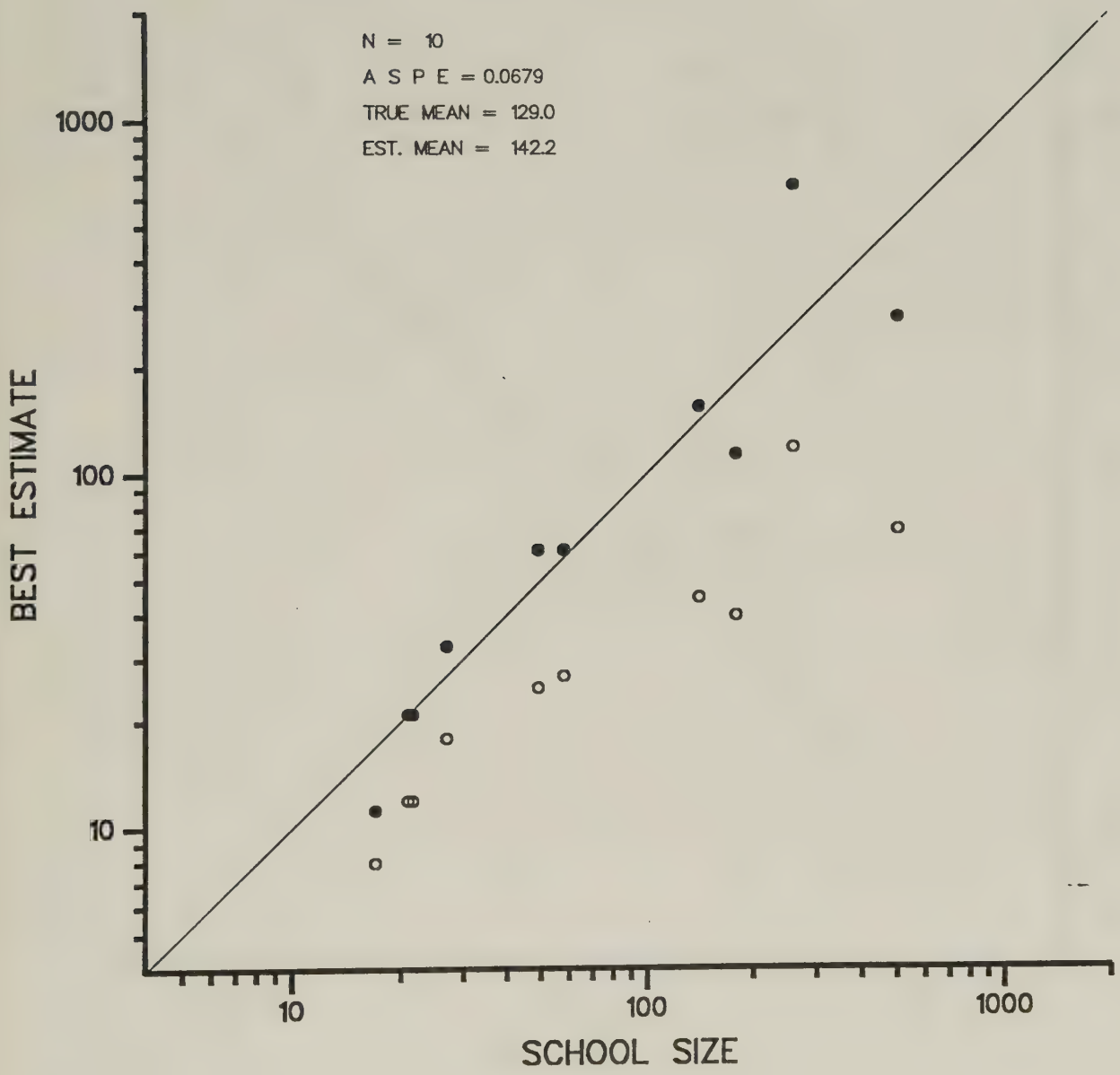


SKOOLWT.FOR
12/18/91

OBSERVER 8

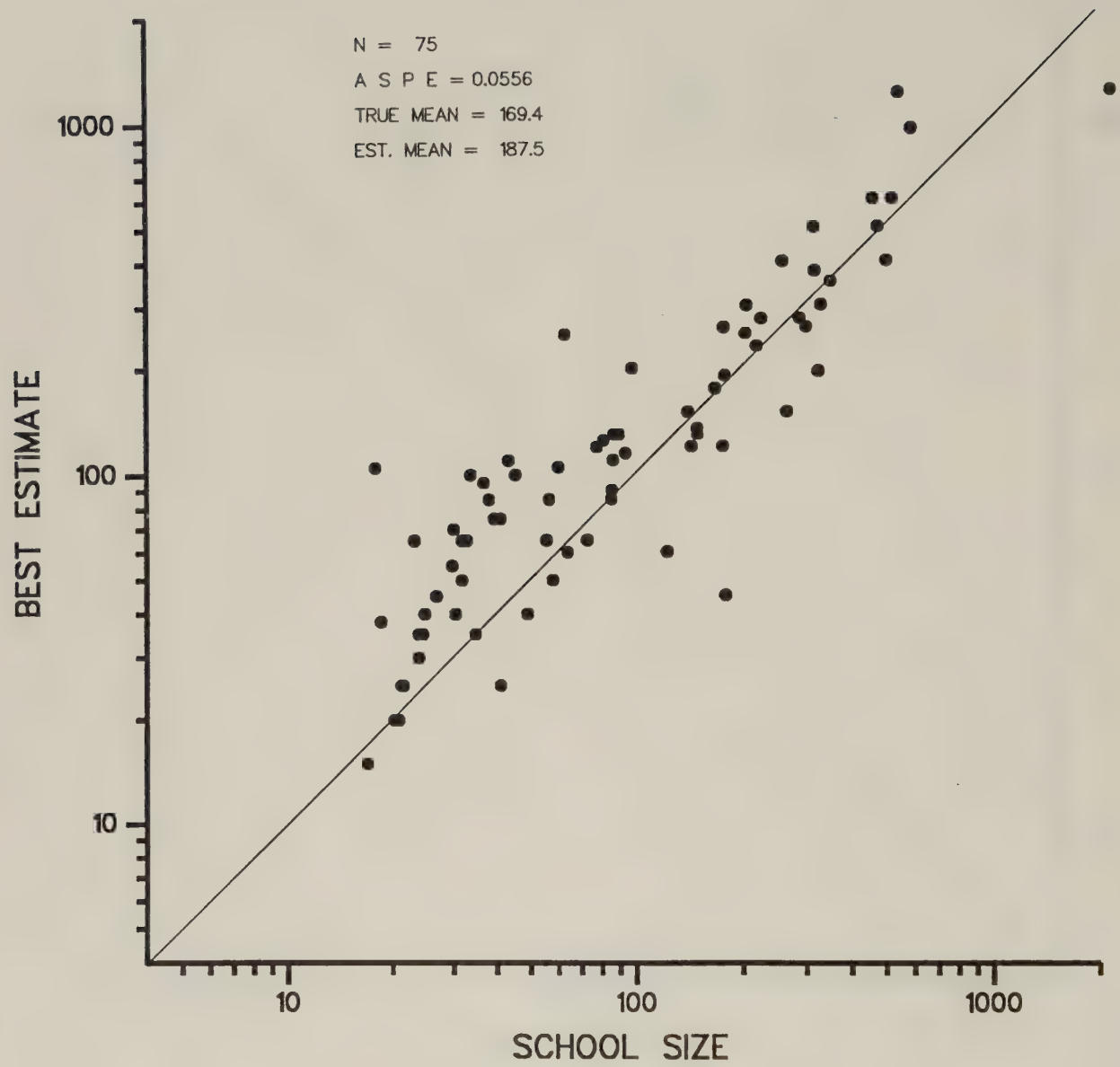


OBSERVER 9


N = 10
A S P E = 0.0679
TRUE MEAN = 129.0
EST. MEAN = 142.2
1000
100
10
BEST ESTIMATE
10
100
1000
SCHOOL SIZE

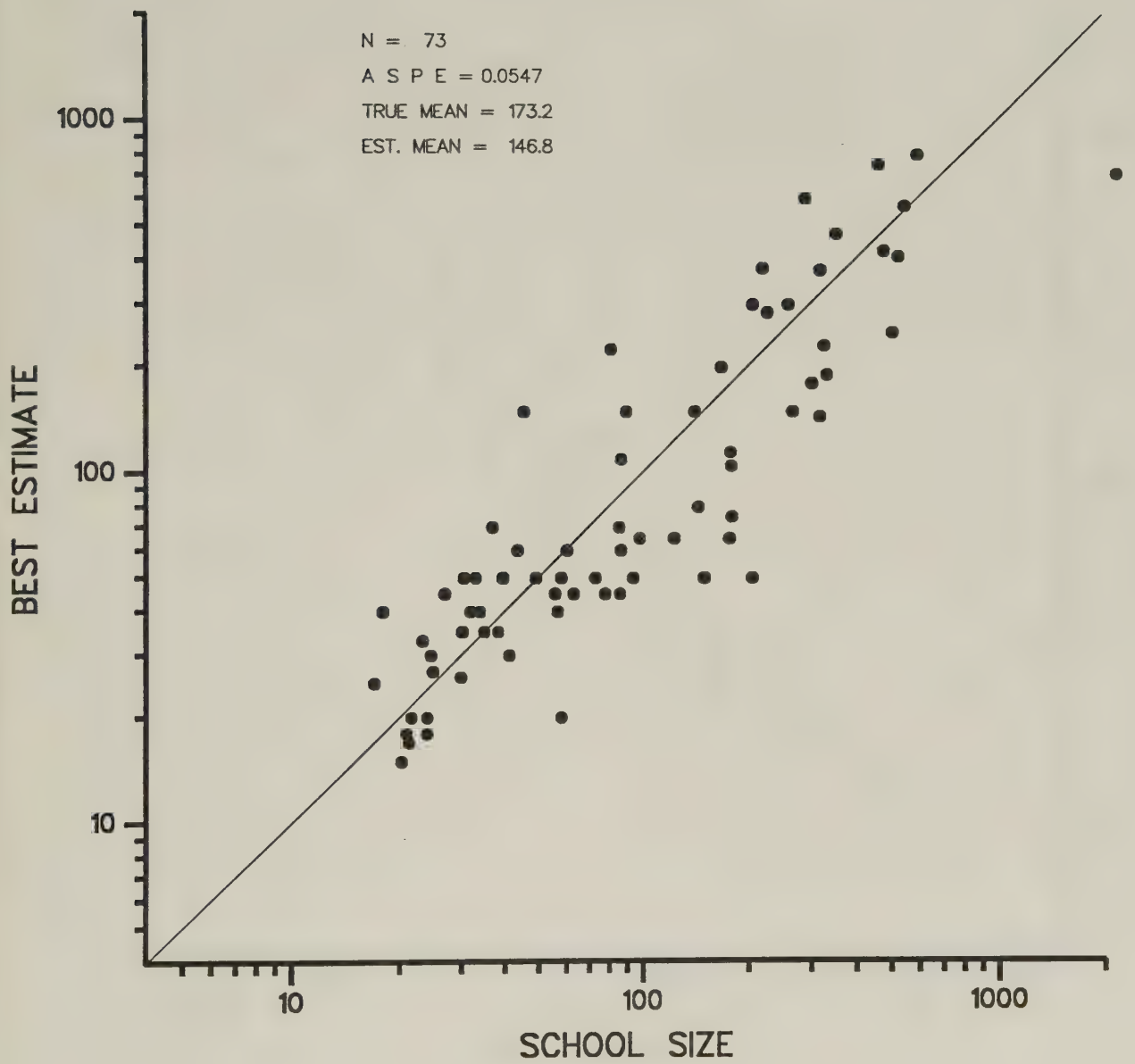
SKOOLWT.FOR
12/18/91

OBSERVER 10



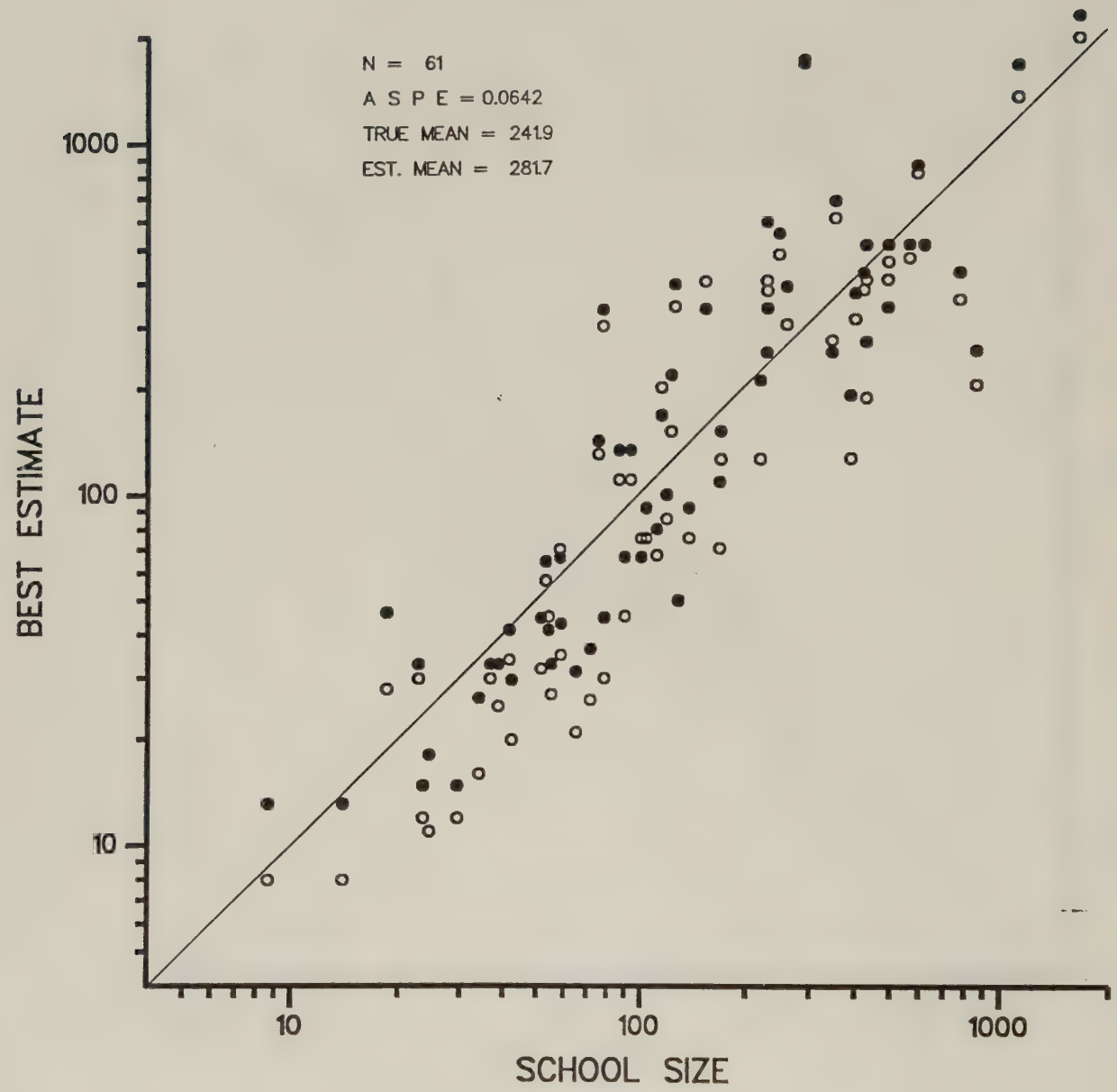
WT.FOR
/91

OBSERVER 11



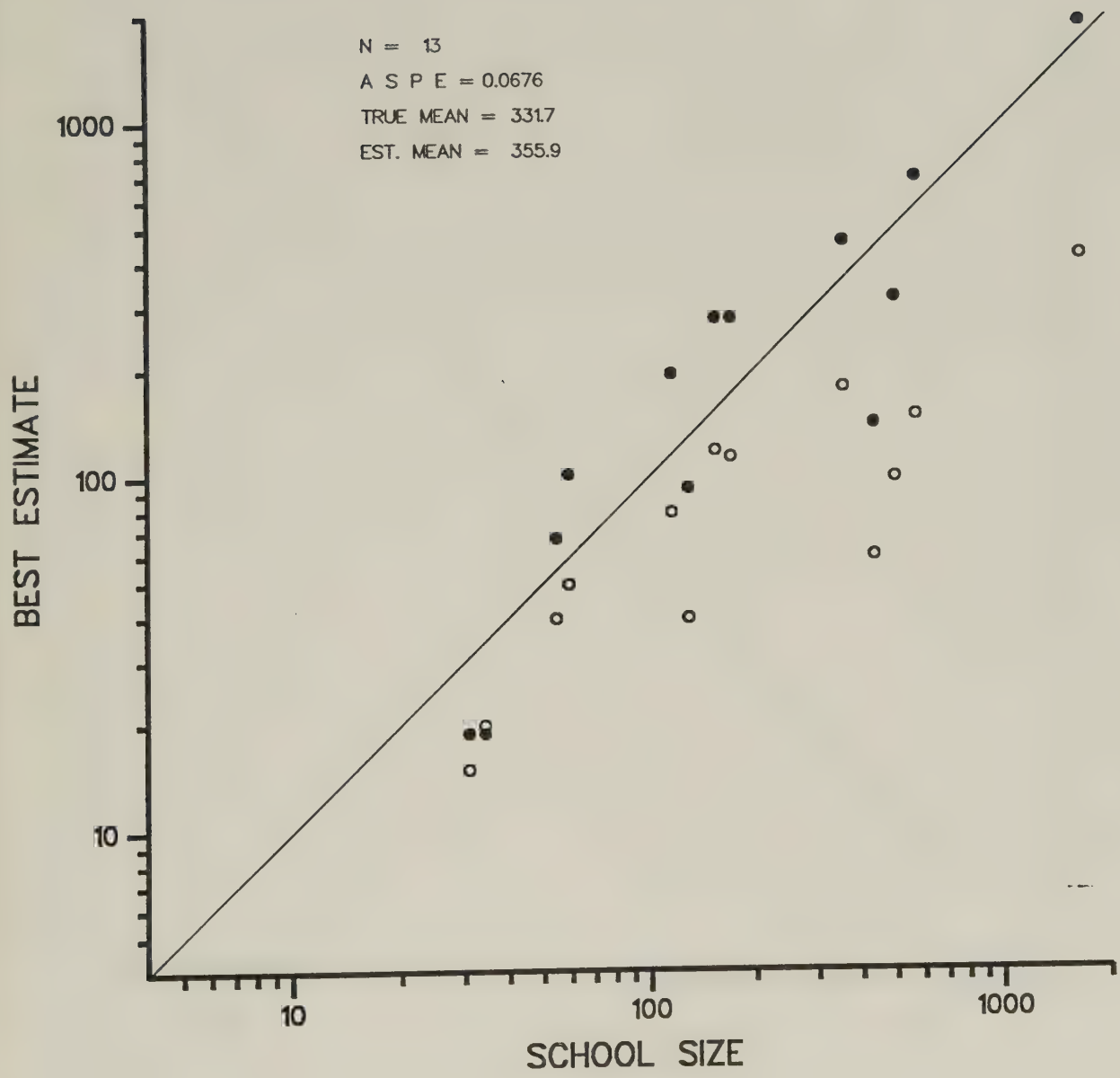
SKOOLWT.FOR
12/18/91

OBSERVER 12

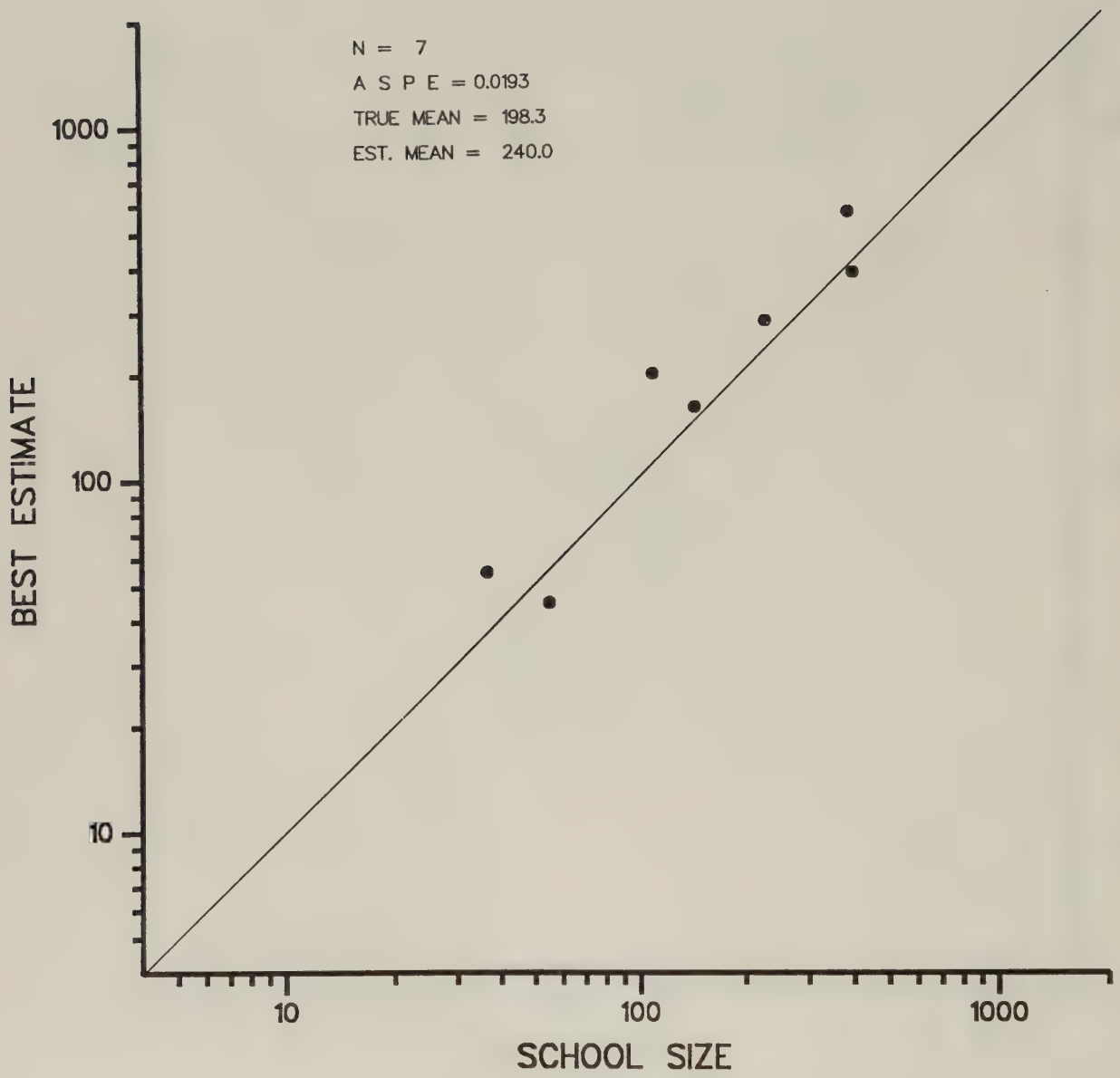


WT.FOR
/91

OBSERVER 13

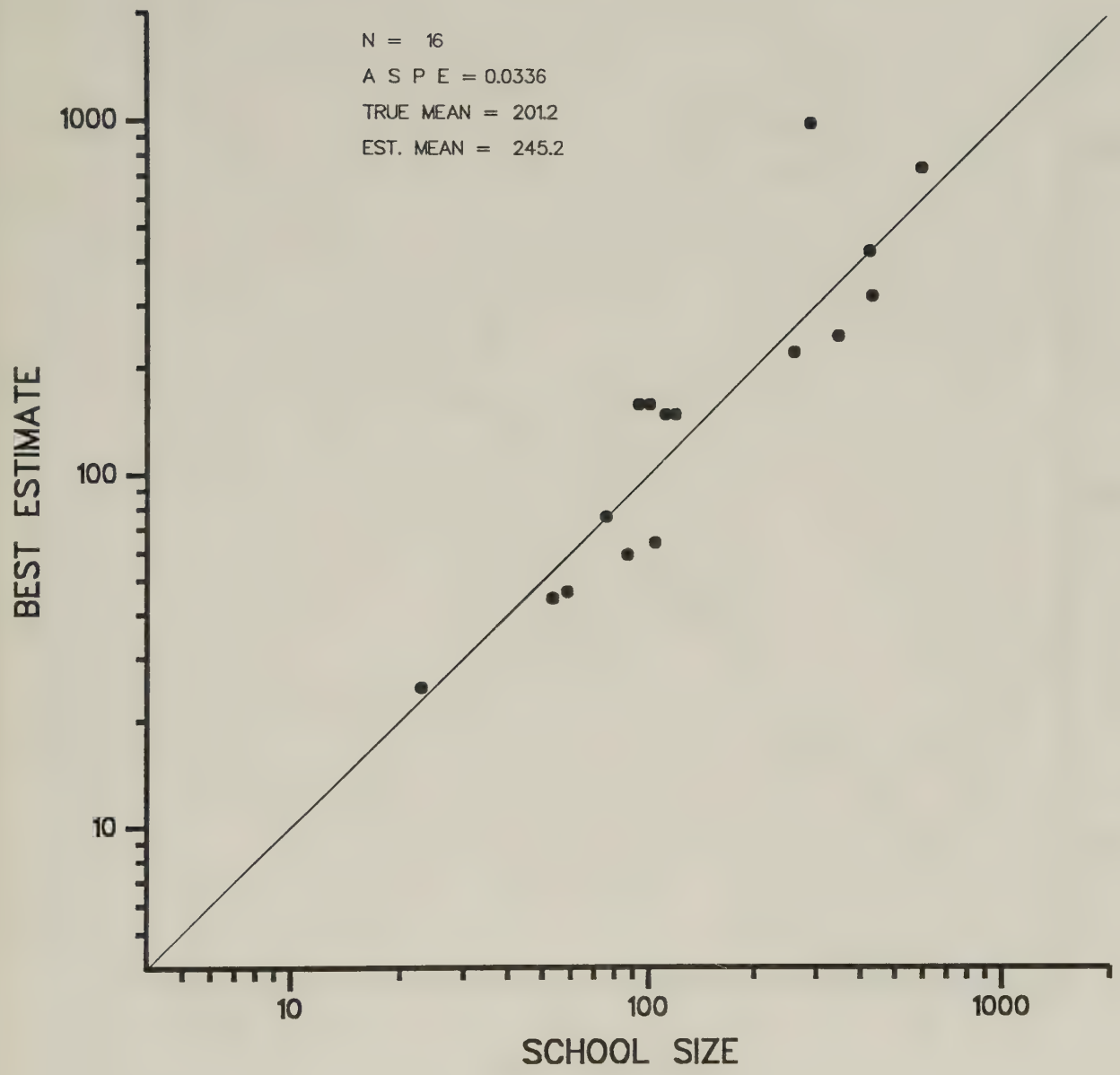


OBSERVER 14

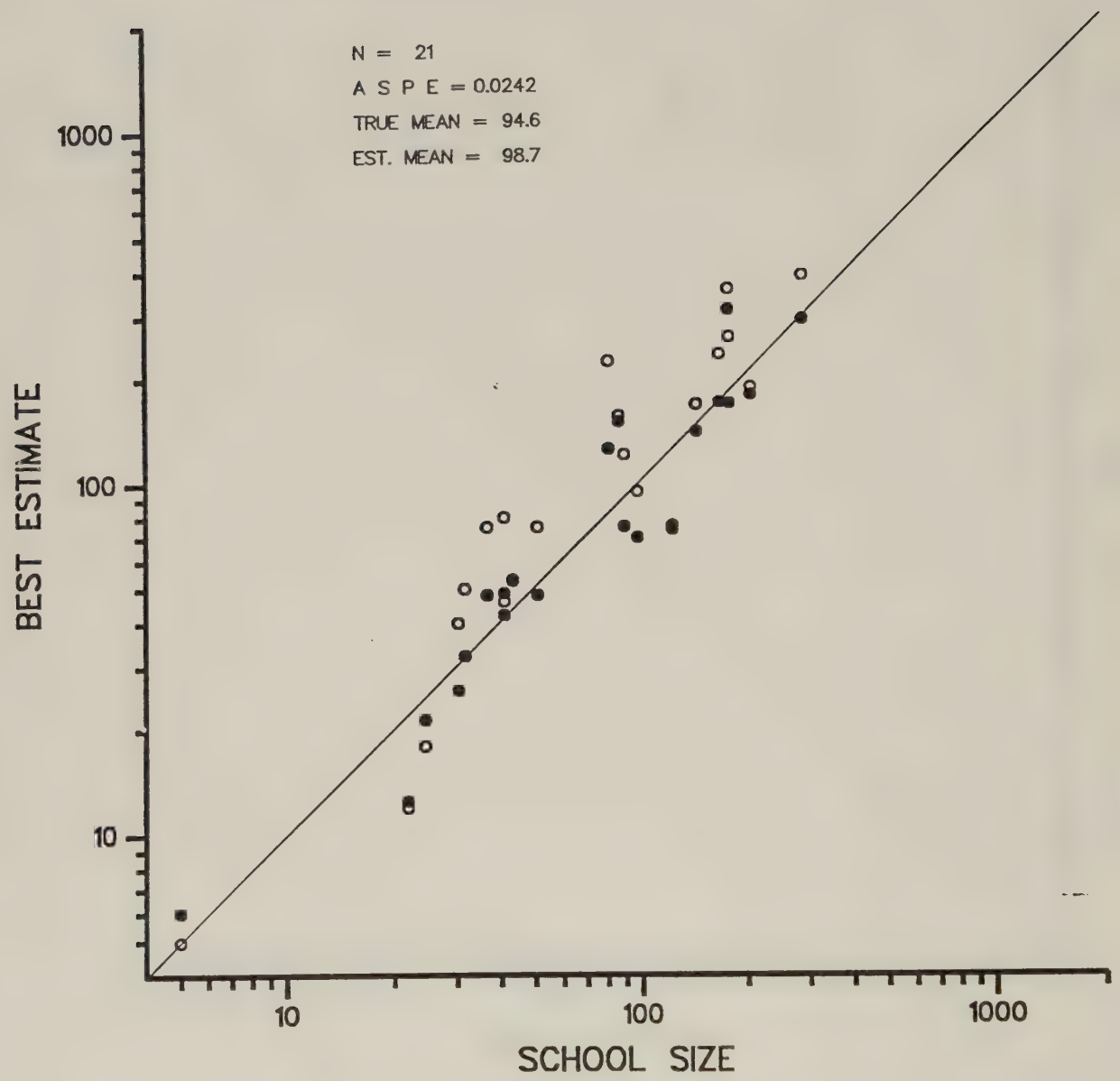

N = 7
A S P E = 0.0193
TRUE MEAN = 198.3
EST. MEAN = 240.0
1000
100
10
BEST ESTIMATE
10
100
1000
SCHOOL SIZE

_WT.FOR
/91

OBSERVER 15

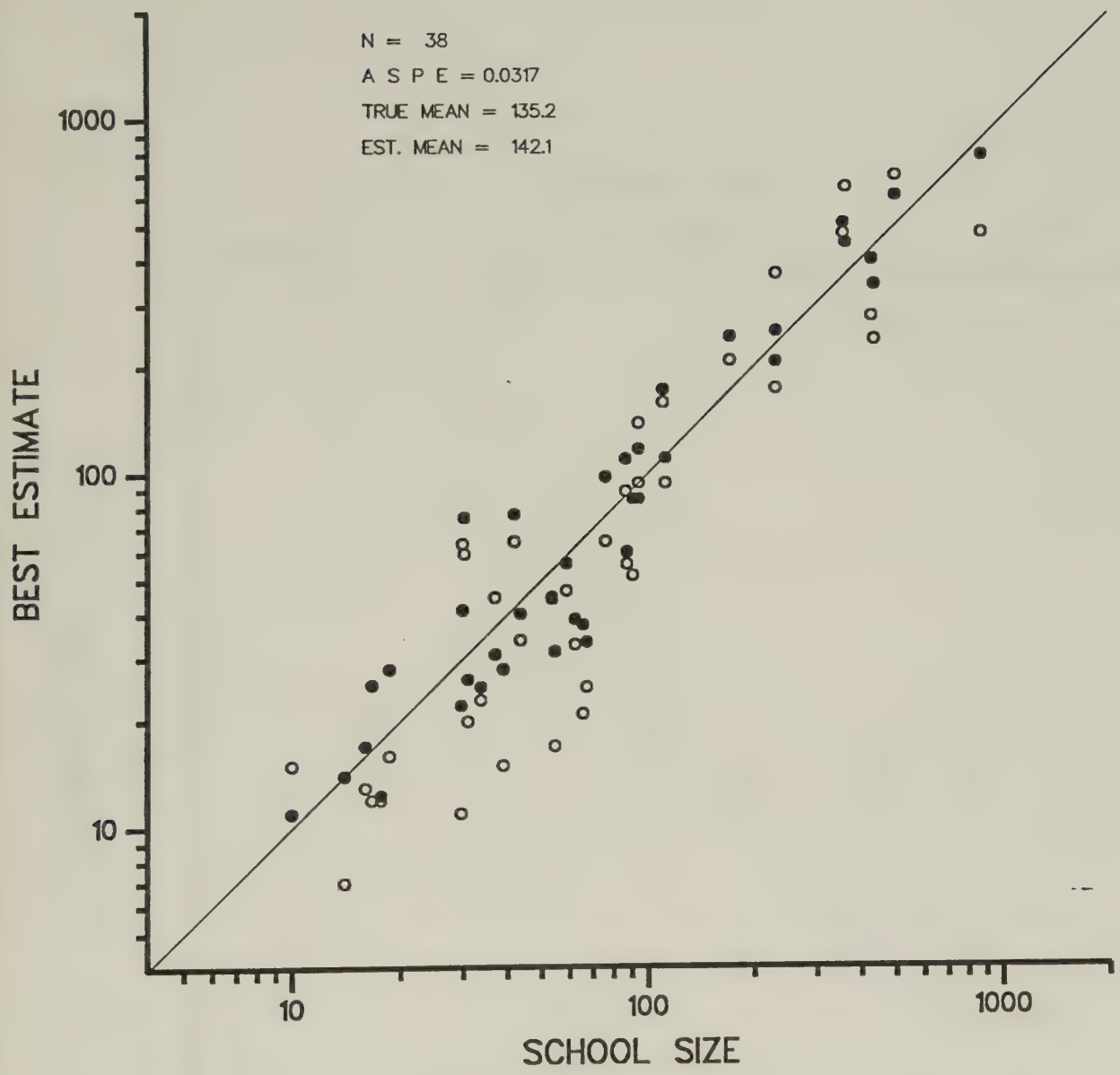


OBSERVER 16

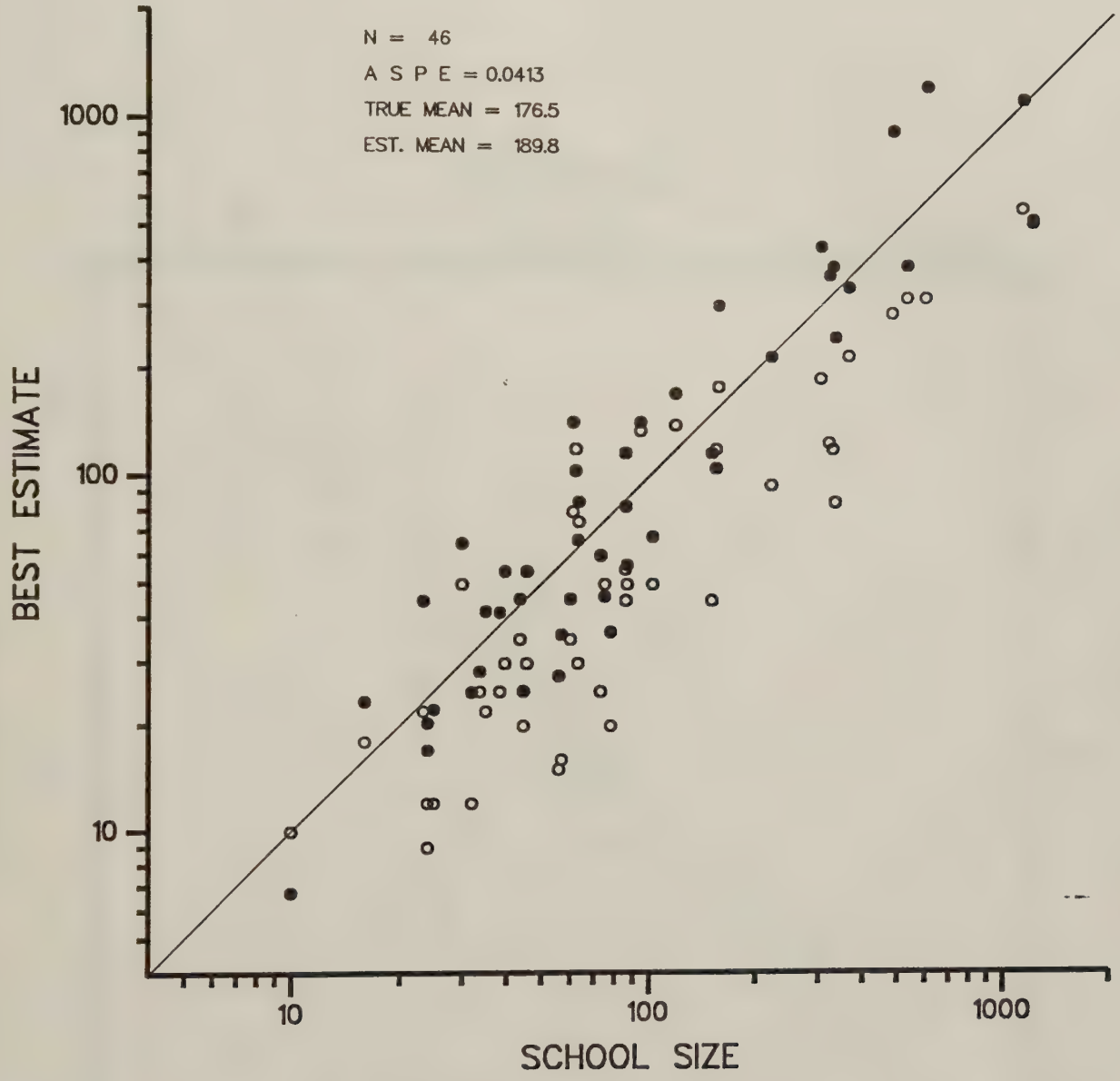

N = 21
A S P E = 0.0242
TRUE MEAN = 94.6
EST. MEAN = 98.7
1000
100
10
BEST ESTIMATE
10
100
1000
SCHOOL SIZE

WT.FOR
'91

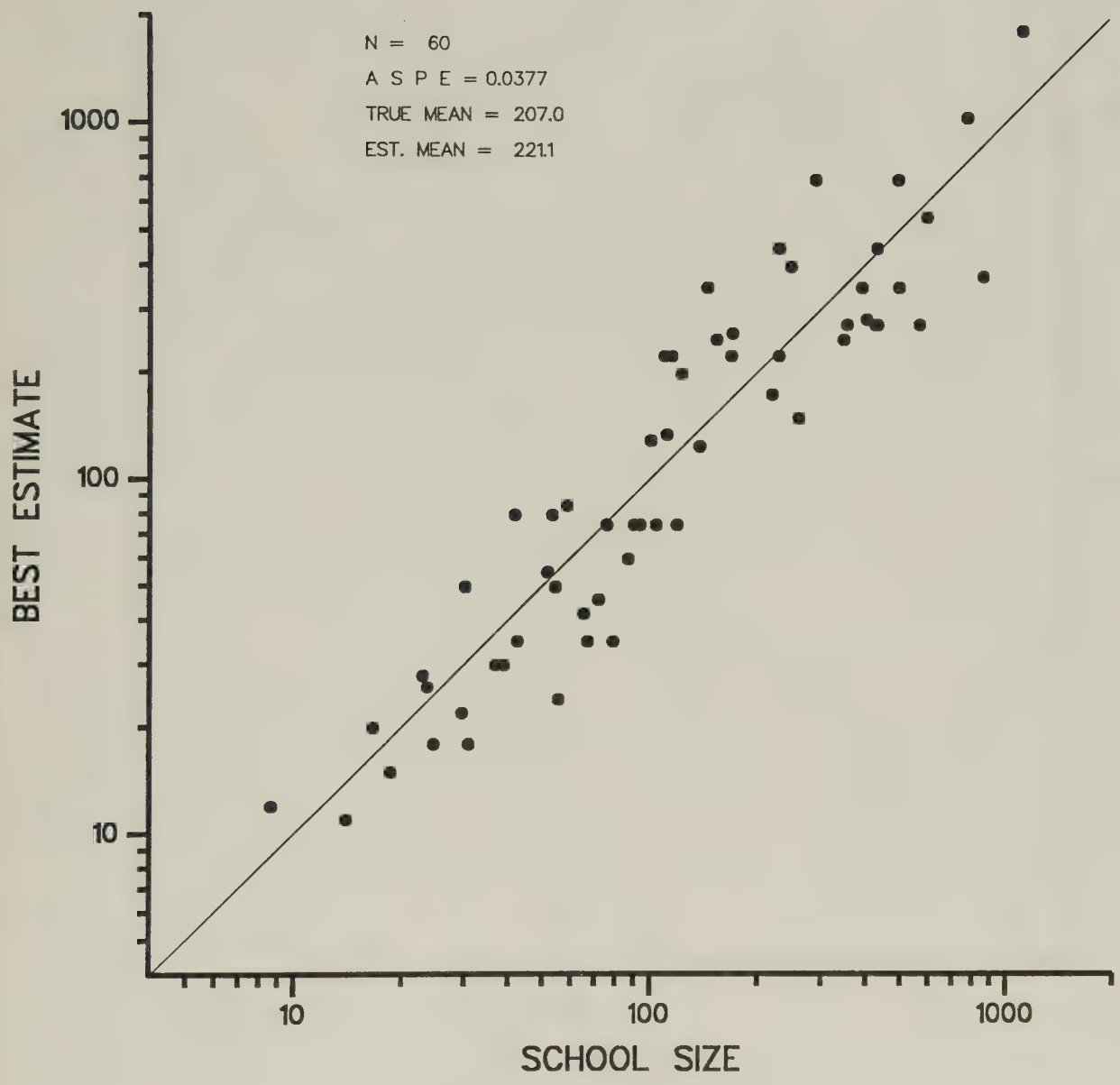
OBSERVER 17



OBSERVER 18

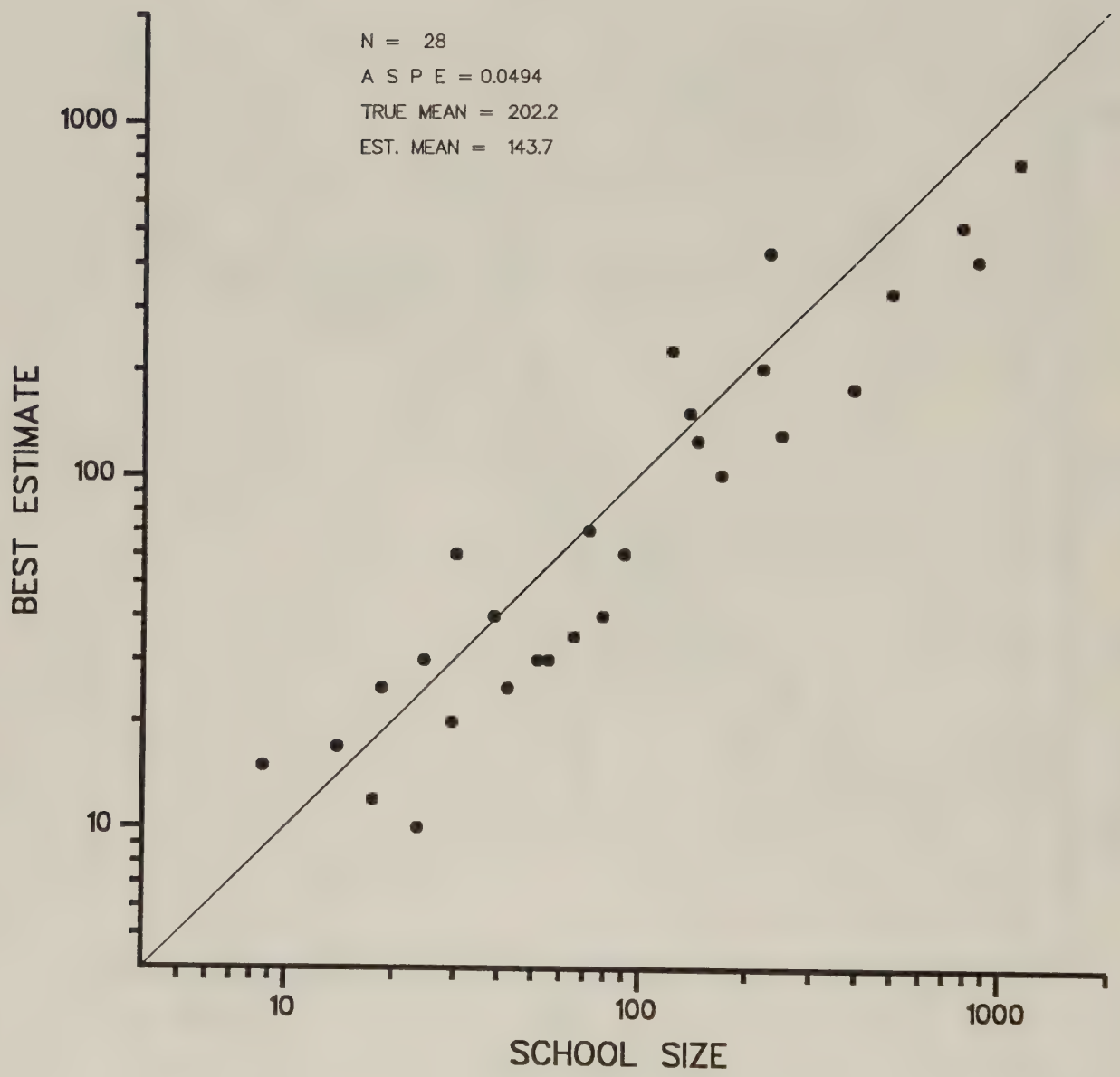

N = 46
A S P E = 0.0413
TRUE MEAN = 176.5
EST. MEAN = 189.8
1000
100
10
BEST ESTIMATE
10
100
1000
SCHOOL SIZE

OBSERVER 19


N = 60
A S P E = 0.0377
TRUE MEAN = 207.0
EST. MEAN = 221.1
1000
100
10
BEST ESTIMATE
10
100
1000
SCHOOL SIZE

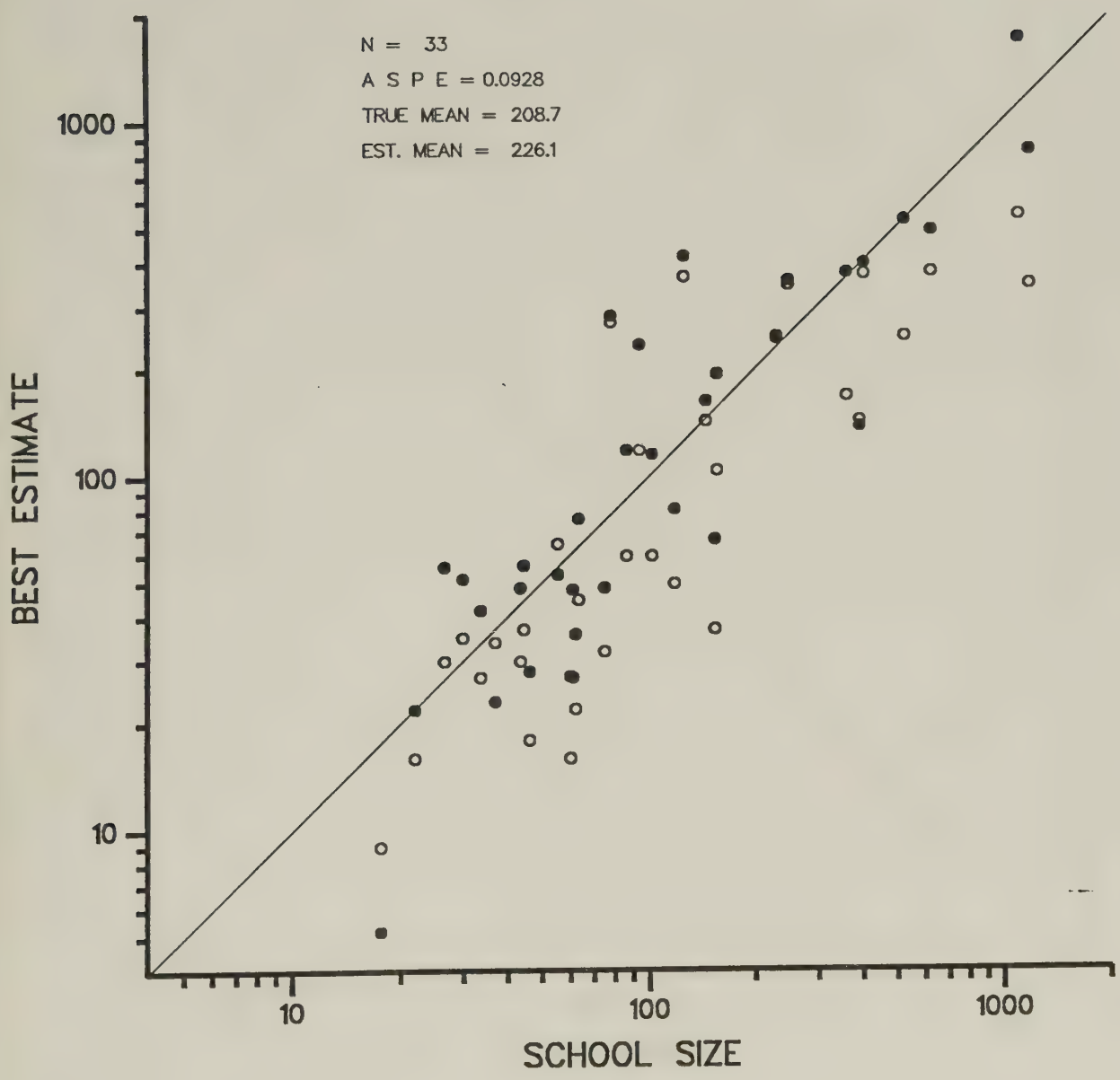
SKOOLWT.FOR
12/18/91

OBSERVER 20

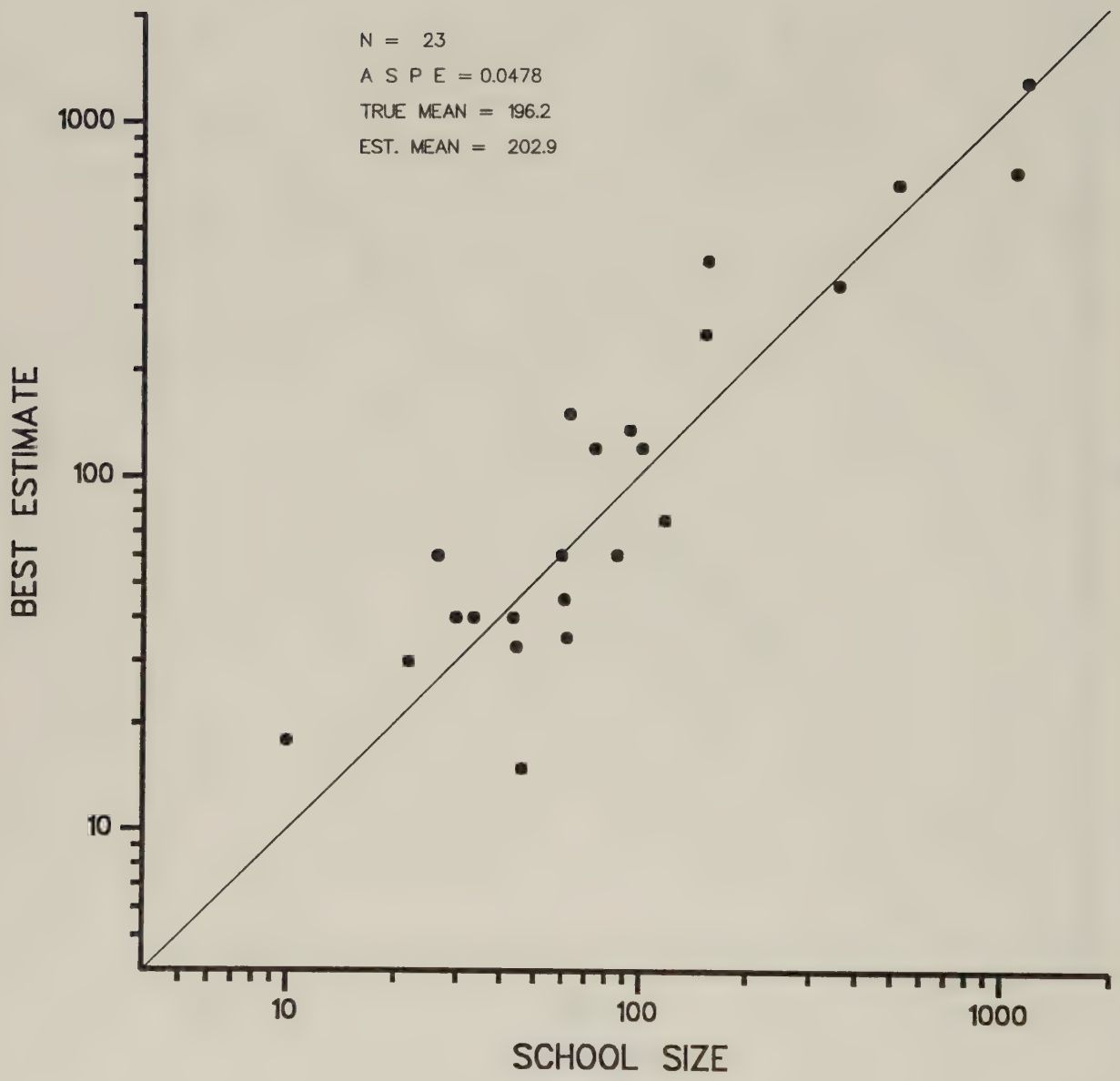


WT.FOR
'91

OBSERVER 21



OBSERVER 22


N = 23
A S P E = 0.0478
TRUE MEAN = 196.2
EST. MEAN = 202.9
1000
100
10
BEST ESTIMATE
10
100
1000
SCHOOL SIZE

WT.FOR
/91

OBSERVER 23

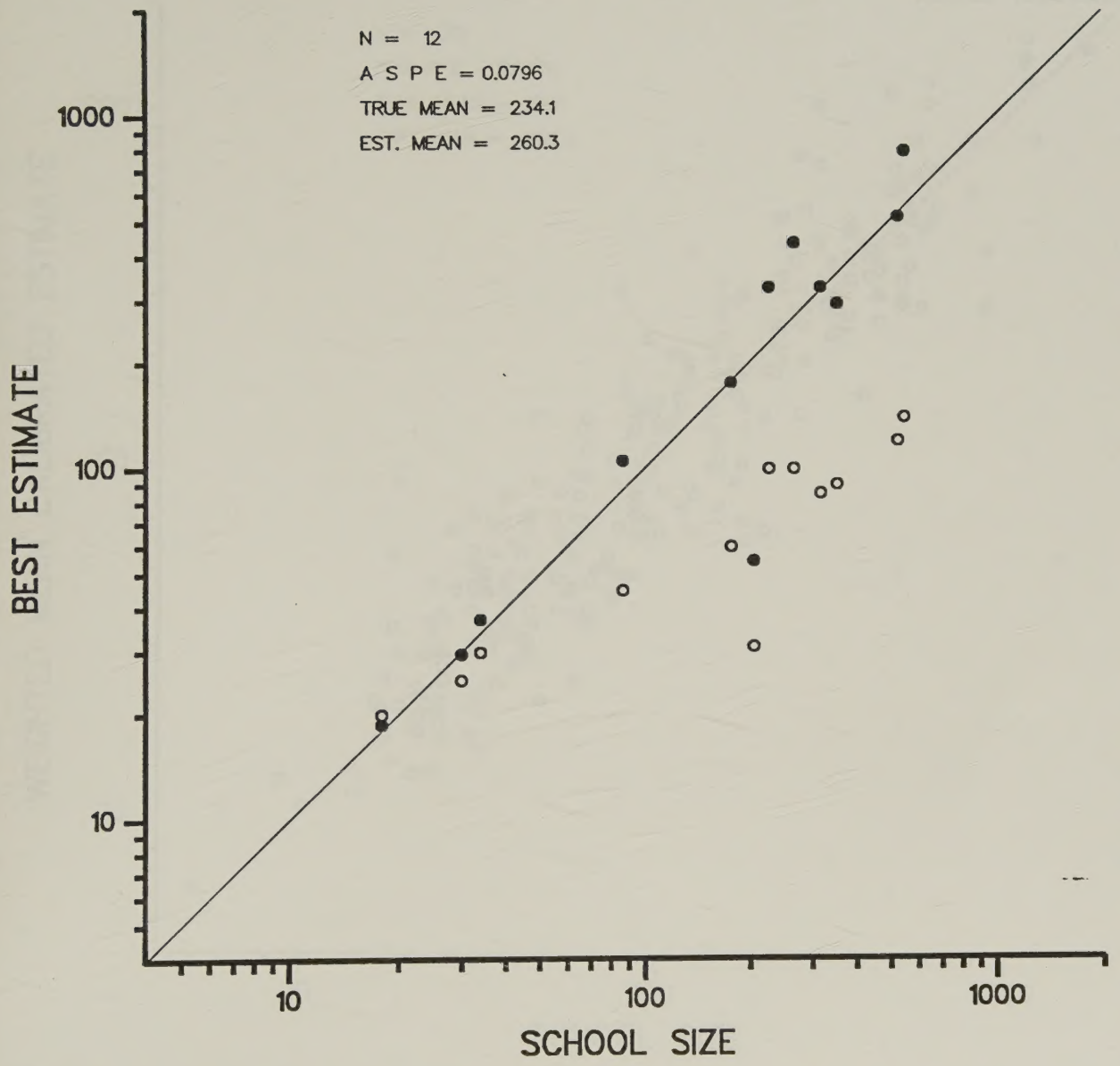


Fig. 3: Weighted mean calibrated estimates of school size plotted against known school size for the entire set of photographed schools (N=316). The diagonal line is the 1:1 line, not a regression line.

-OL.FOR
19/91

WEIGHTED MEAN CALIBRATED ESTIMATES